DOCUMENT RESUME

ED 187 740 TM 800 236

AUTHOR	Scott, Laurie
TITLE	Summary of the Fall 1978 Conference of the National Consortium on Testing: October 23-24, 1978.
INSTITUTION	Huron Inst., Cambridge, Mass.
PUB DATE	9 Nov 78
NOTE	91p.; For related documents, see TM 800 235 and TM 800 237.
AVAILABLE FROM	Huron Institute, National Consortium on Testing, 123 Mt. Auburn Street, Cambridge, MA 02138 ($5.00)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; *Educational Assessment; *Educational Testing; Evaluation Methods; Minimum Competency Testing; Professional Associations; Standardized Tests; Standards; State Legislation; Testing Problems; *Test Reviews; Test Selection
IDENTIFIERS	Alternatives to Standardized Testing; California; *National Consortium on Testing; Project TORQUE

ABSTRACT

Discussions at the National Consortium on Testing conference are summarized. Updates on recent and upcoming events are given by Jeff Schiller (National Institute of Education), Doug Whitney (American Council on Education), John Maxwell (National Council of Teachers of English), Dorianne Almann (National Science Teachers Association), Eric Gardner (National Council on Measurement in Education), and Walt Haney (Huron Institute): Ron Hambleton's presentation on the state of criterion-referenced testing and Banesh Hoffman's remarks on declining writing ability are summarized. Alternative approaches to educational assessment are discussed by Vito Perrone, George Hein, Fritz Mosher, and Judah Schwartz, who describes Project TORQUE. Topical sessions on minimum competency testing and standards for tests and test use are led by George Madaus and Walt Haney, respectively. Deborah Coates also discusses the National Academy of Science's Committee on Ability Testing. Appendices include a summary of California Testing legislation, guidelines for evaluating criterion-referenced tests and test manuals, bibliographies, standards for test use, newsletters/announcements, and an outline for reviews of commonly used achievement and diagnostic test series. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187740

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SUMMARY OF THE FALL 1978 CONFERENCE

OF THE

NATIONAL CONSORTIUM ON TESTING

(October 23-24, 1978)

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Paul L. Houts
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Prepared by

Laurie Scott

Huron Institute
123 Mount Auburn Street
Cambridge, Mass. 02138

November 9, 1978

TM800 236

This summary has been prepared from notes and tape recordings of the substantive meetings of the conference. It does not represent a verbatim account of the conference.

TABLE OF CONTENTS

		Page
I.	Introduction	1
	Vito Perrone	1
	David K. Cohen	1
II.	Update: Recent and Upcoming Events Concerning Testing	3
	Jeff Schiller	3
	Doug Whitney	6
	John Maxwell	7
	Dorianne Almann	8
	Eric Gardner	9
	Walt Haney	10
III.	The State of the Art of Criterion-Referenced Testing	10
	Ron Hambleton	10
	Discussion	13
IV.	Remarks by Banesh Hoffman	16
V.	Alternative Approaches to Educational Assessment: A Panel	17
	Vito Perrone	17
	George Hein	17
	Fritz Mosher	19
	Discussion	21
VI.	Alternative Approaches to Educational Assessment: An Example	22
	Judah Schwartz, Project Torque	22
VII.	Topical Sessions	27
	Introduction	27
	Alternative Approaches to Educational Assessment; Vito Perrone	27
	Minimum Competency Testing; George Madaus	28
	Standards for Tests and Test Use; Walt Haney	30
VIII.	Remarks by Deborah Coates	31
IX.	Report on Huron's Reviews of Major Standardized Test Series	32
	Walt Haney	32
	Laurie Scott	33
	Discussion	33
X.	Footnotes	36

Page

XI. Attachments 39

Summary of California law to regulate activities of testing services 39

Jeff Schiller, "The Testing, Assessment and Evaluation Program" 41

Doug Whitney, "A Brief Description of the American Council on Education Credit-by-Examination Test Review Program" 49

"SLATE," Vol. 2, No. 2, December 1976 55

Ronald K. Hambleton and Daniel Eignor, "Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals" 57

Ronald K. Hambleton, Criterion-Referenced Testing Bibliography 75

George E. Hein, "Alternative Approaches to Educational Assessment" (bibliography) 77

APGA Policy Statement, "Responsibilities of Users of Standardized Tests" 79

Committee on Ability Testing Announcement 87

Committee on Ability Testing, description and list of Committee members and staff 89

Outline for Descriptive Review of Common Standardized Test Series 91

SUMMARY OF THE FALL 1978 CONFERENCE

OF THE NATIONAL CONSORTIUM ON TESTING

I. Introduction

The Fall 1978 Membership Conference of the National Consortium on Testing was held October 23 and 24 at the Sheraton-Park Hotel in Washington, D. C. This paper summarizes the substantive meetings on Monday and on Tuesday mornings.

Consortium President Vito Perrone opened the meeting Monday morning by welcoming to the conference Consortium representatives and other guests. He commented that the Consortium was organized to examine a broad range of testing issues which relate principally to schools but also affect society more broadly. That examination has begun and more information is becoming available in part because of member organizations' own activities.

Vito noted that competency testing continues to receive public attention and Consortium concern. This concern, he feels, is rooted in the NCT's basic commitments to public education, quality schooling, and equity. For these reasons, it remains important for the Consortium to stimulate public debate and related public policy analyses. The Consortium, he said, has the potential to provide some educational leadership and to reject the despair over apparent declines in competence.

Vito pointed out that the plans for Monday's meetings were based on Consortium interests and remain very close to the plans for the first year of the Carnegie grant.

David K. Cohen, Executive Director of the Consortium, presented an update on the Huron Institute's staff work on behalf of the Consortium. David

noted that Huron has received many requests for information about the NCT. He added that he and Walt Haney have been asked to speak at a number of member organizations' meetings.

David then discussed the staff work being done in the four program areas originally outlined: truth in testing, the federal government's role in testing, standardized testing at the state and local levels, and alternative approaches to assessment. In regard to truth in testing, several staff circulars have been prepared which deal with the quality of standardized tests. David drew attention to the discussion of Standards for Tests and Test Use scheduled for Tuesday morning. He noted that reviews of several most commonly used standardized tests were being done and that a staff circular based on the reviews would be available within a few months.

The second area of Huron's work, David continued, is the federal government's role in testing. He called attention to Jeff Schiller's presentation following his own report. David remarked that he and Walt Haney had participated earlier this year in the HEW Conference on Basic Skills and Achievement Testing. Huron's work, he explained, would focus primarily on federal requirements for program evaluation and testing in the area of special education. Staff circulars on these issues, he advised, would be available in five or six months.

The third area of concern is the role of standardized tests at the state and local levels. Most of Huron's efforts in this realm, he reported, have focused on minimum competency testing. Walt Haney and he presented a paper on minimum competency and social policy at a recent AERA meeting; this paper will be revised as a staff circular. Also, a staff circular on minimum

competency by Walt and George Madaus has been revised and will appear in the next issue of the Harvard Educational Review.

David noted too that a topical session on minimum competency testing would be held on Tuesday morning.

In addition, Elinor Woods of the Huron Institute is preparing an analysis of the evolution of state testing and assessment programs and their impact on education. She expects to have a staff circular ready in the relatively near future. Huron hopes to extend this analysis to testing programs in large cities, David reported.

He further reported that he had been in California talking to the staff of the Senate Education Committee which has proposed legislation regarding the use of tests statewide. The bill was signed into law the week before the NCT conference and is the first attempt in the history of the country, David stressed, to try to regulate the use of tests.[1]

Finally, in the area of alternative assessment strategies, David observed that Huron is collecting information on alternatives and hopes to produce several staff circulars on alternative assessment procedures in the second year of the Carnegie grant.

II. Update: Recent and Upcoming Events Concerning Testing

The first update on events concerning testing came from Jeff Schiller, Director of the NIE Office of Testing, Assessment and Evaluation. He discussed that office's present and future plans. He first described the reorganization of NIE into three major program groups: (1) dissemination, (2) educational policy and organization, and (3) teaching and learning. The third

group is headed by Sylvia Scribner, and the new NIE Office of Testing, Assessment and Evaluation is located in this group. Jeff noted the NIE has spent relatively little money on testing and assessment. Right now they are in the middle of a six-to-eight month definitional phase to decide where to put their money. In three or four months they expect to have a five-year plan for funding for 1979-1983.

Jeff went on to outline the three major strategies being pursued in the area of testing: improving current testing practices, clarifying the role and impact of testing in education, and promoting the integration of testing with instruction and learning.[2] He offered several examples of funding in these areas. In the area of improving current testing practice, NIE is working on the development of a pamphlet on the effects of standardized achievement and intelligence tests for the Government Services Administration (GSA). He explained that two and a half years ago the GSA did a survey to determine public demand for brochures on various topics. The number one concern revealed in the survey was the effects of testing, surpassing the previous issue of greatest interest, the removal of stains from fabric. Demand for a pamphlet on testing is estimated to be between 150 and 170 thousand copies. Also, Jeff reported that an award would be made to the American Federation of Teachers and the Huron Institute to develop materials on testing for teachers and to hold workshops. NIE also is planning a series of regional workshops, to be held between January and June 1979.

In the area of clarifying the role and impact of testing in education, NIE is developing a summary of the status of minimum competency testing, from a neutral standpoint. In this area Jeff also announced that an award

had been made to Judah Schwartz of M.I.T. to explore the social context of testing.

Jeff discussed briefly an NIE conference on testing and technology held at Falmouth, Massachusetts. He observed that the conference seemed to suggest the emergence of a new discipline consisting of a mix between the cognitive psychologists and the computer scientists. The conference was part of NIE's strategy to investigate the integration of testing with learning.

Three strategies are also being pursued at NIE in the area of evaluation, Jeff observed. These are the identification and evaluation of promising field programs, advancement of the art of evaluation methodology, and the strengthening of the capacity of federal, state, and local groups to conduct evaluations.[3]

Jeff concluded by outlining four methods NIE uses to dispense money:

1) Targeted research through Requests for Proposals (RFPs);

2) Unsolicited proposal competition;

3) A teaching and learning grants competition to be announced in November. Bidders will have two and a half to three months to respond. Awards will be made in the late spring; and

4) Sole source awards for organizations which are doing research and need additional funds to do another piece of the research, for example, to conduct an evaluation of a demonstration project.

Several questions and comments were made after Jeff's presentation. Walt Haney raised the point that, although the NIE-funded CSE Test Evaluations led to the conclusion that about 90% of the tests evaluated do not meet the CSE criteria of adequacy for program evaluation, Jeff still talked a great deal about testing in regard to evaluation. Jeff responded that this would

in fact be taken into account by NIE as it develops policies in the area of current testing practices.

Lillian Weber commented that she did not hear in Jeff's talk, any support for studying how teachers make judgements, about how much they in fact use the information from the tests, and what they use when they are not using these external tests. She suggested that this could be the basis of a proposal to NIE.

Banesh Hoffman advised that there are two types of evidence, statistical evidence and "intangibles," including anecdotal and experiential documentation. He expressed concern that the intangible evidence might be ignored because it is not "scientific," and he wished to prevent such a mistake. Jeff remarked that not only would such non-statistical evidence be considered by NIE, but that consideration of the ways in which qualitative evidence can be used is likely to be part of the grants competition in the following year.

The second presentation in the update session was by Doug Whitney, Associate Director of the Office on Educational Credit of the American Council on Education. He presented some background information on the Council's activities regarding the award of high school and college credit for skills acquired outside the classroom. He summarized four current Council concerns about the credit-by-examination policies of post-secondary institutions. He went on to detail the Council's test review program.[4] Doug also drew attention to two papers, "Guidelines for Reviewing the Content of Examinations for Granting College Credit" and "Guidelines for Reviewing the Technical Quality of Examinations for Granting College Credit," which can be obtained by writing to:

American Council on Education
Office on Educational Credit
Suite 1
One Dupont Circle, N. W.
Washington, D. C. 20036

Discussion followed Doug's presentation. Paul Jacobs, from Thomas A. Edison College and author of Up the IQ, inquired about the sources of tests reviewed by the Council. Doug replied that test publishers were asked to submit tests voluntarily. He added that the Consortium might aid the Council in identifying tests for review by submitting recommendations. William Schafer asked which national organizations provided technical assistance to the Council. Doug answered that the NCME and AERA advised the Council, and he urged NCT members to suggest other national and regional organizations which might be of assistance. Walt Haney observed that the tests considered by the Council might be thought of as the converse of minimum competency testing, since they are used to award rather than to deny credit. He asked if the Council had considered looking into other sorts of tests and extending their validation work. Doug responded by identifying the scope of the Council's interest. The Council assesses tests used by more than one institution for the awarding of post-secondary credit. The Council does not, he explained, look into elementary or secondary tests, licensing exams, or tests for permits to practice. However, he observed that if minimum competency testing is extended to the college level, for example to determine eligibility to receive the baccalaureate, that area of MCT would be of concern to the Council.

Continuing the update, John Maxwell of the National Council of Teachers of English, described the formation of the ad hoc committee entitled SLATE (Support for Learning and Teaching of English) within the NCTE. He explained

that the group is a political one, initially formed to see if the organization could have some effect on those who make decisions affecting the teaching of English. Since the group could not take on everything, it decided to focus on testing and evaluation as their first issue of concern. The result has been a series of SLATE newsletters or "Starter Sheets" designed as a means of educating their own membership and informing legislators of issues SLATE feels are important.[5]

Dorianne Almann of the National Science Teachers Association expressed concern on the part of science teachers about the minimum competency testing movement's effect on science instruction. She reported the results of a survey conducted by the NSTA of 47 state science supervisors. Of the 80% who responded, 93% reported that science is not assessed in statewide minimum competency testing programs. Seven percent (Maryland, West Virginia, and Washington) responded that science is included in the statewide assessment programs. Only five percent of the respondents stated that lists of science skills taught are available.

Dorianne noted the apprehension of the NSTA that science, which now is often taught as an afterthought at the elementary school level, might receive even less attention as the minimum competency movement grows. She also stressed that the MCT movement may lead to a shift in emphasis from process to content in the science curriculum.

Following Dorianne's presentation, Lillian Weber emphasized that science and social studies are indeed often "put off until after Christmas" in classroom teaching. She commented that the lack of questions regarding Dorianne's talk reflected support for her point of view from the Consortium

members. Vito Perrone added that "basics" need to be redefined much more broadly to encompass more than just reading and mathematics.

Eric Gardner, representing the National Council on Measurement in Education, provided an update on four NCME activities. First, a joint committee of the NCME, APA, and AERA has reexamined the Standards for Educational and Psychological Tests and has decided that a revision is needed. The committee has stressed that the predominantly technical standards need to focus more on test users' problems.

Second, three years ago an NCME committee chaired by Daniel Stufflebeam concluded that a set of standards, comparable to those for testing, should be developed for the field of evaluation. A first draft of these standards has been completed and circulated for comment. It is anticipated that a final version will be available for use in about a year.

Third, it is well-known that NCME publishes the Journal of Educational Measurement and, Eric observed, that the journal tends to be written for technical people. Another NCME journal, Measurement in Education, is published four times a year and is designed primarily for guidance counselors, teachers, and administrators. Herb Rudman has just become the editor of this journal and is planning a series of issues for the next three years which will be more useful to school boards, parents, and practitioners.

Fourth, an NCME Committee has been set up to examine the problems of dissemination, misuse, and use of tests. The Committee has no specific project as yet, because it has wanted to avoid duplication of efforts. George Madaus, Eric related, has submitted a proposal through the AFT to be funded by NIE; the Committee will try to augment George's proposal.

Walt Haney of the Huron Institute presented an update on recent and future events of concern to the Consortium. He noted that Daniel Stufflebeam, chairman of the NCME committee which has developed a set of draft standards for program evaluation, has approached the NCT about holding hearings on the standards in Washington in conjunction with NCT's spring meeting. Walt promised to keep Consortium members informed about this.

Another issue of interest to Consortium members, Walt observed, is the first federal court challenge to competency testing, filed against the state of Florida by ten black high school students who failed the state's minimum competency exam. The students allege that the test is statistically biased against blacks, that it violates the 14th Amendment due process and equal protection clauses and the Civil Rights Act's Title VI race-discrimination ban, that cutoff scores were set arbitrarily, and that items were not examined for reliability and validity. The students are requesting a permanent injunction against the use of the test as a prerequisite for graduation.[6]

Walt also brought to Consortium attention a study conducted by the Federal Trade Commission into test coaching schools, specifically the possibility of misleading advertising. Walt continued that he tried to have an FTC representative at the Consortium meeting to discuss the findings, but could not since the study is presently undergoing review. Vito Perrone added that there was definitely Consortium interest in following up on the FTC report.

III. The State of the Art of Criterion-Referenced Testing

The principal speaker Monday morning was Ron Hambleton of the University

of Massachusetts at Amherst, who spoke to the Consortium on "The State of the Art of Criterion-Referenced Testing." Ron noted that two seminal papers were responsible for getting people to think about criterion-referenced testing; one by Glaser in American Psychologist, 1963; another by Popham and Husick in the Journal of Educational Measurement, 1969. Since then, he noted, 700 papers have been written on the topic, but about 650 of them can be put aside because there is growing consensus of what criterion-referenced testing technology should look like.[7]

Ron then defined the terms criterion-referenced testing (CRT), objectives-referenced testing (ORT), domain-referenced testing (DRT), competency testing (CT), and minimum competency testing (MCT). He pointed out that strictly speaking, a criterion-referenced test is a test designed to determine an examinee's level of performance relative to a well-defined behavior domain. Thus the term would encompass competency tests and domain-referenced tests. An objectives-referenced test, on the other hand, does not require that an hypothesized pool of items or domain is specified.

Ron went on to compare CRT with norm-referenced testing (NRT). He described three differences which distinguish norm- and criterion-referenced testing. First, whereas CRT must start with clearly defined domain specifications, NRT writes items to fit a set of topics and then chooses specific items to meet certain statistical qualifications. Second, the two types of tests differ in the way items are selected. NRT chooses items which will maximize discrimination among examinees; CRT emphasizes the match between items and the domain. And last, the testing strategies differ in intended use. NRT is designed to compare one student with another, but

CRT is intended to facilitate interpretations about content mastery.

Since CRT and NRT are different strategies, they require different technologies, Ron advised. He listed twelve steps in the development of CRT:

1) Preparation and/or Selection of Objectives
2) Preparation of Test Specifications
3) Writing Test Items
4) Editing Test Items
5) Determining Content Validity
   a) Content Specialists
   b) Item Analyses
6) Further Editing of Items
7) Test Assembly
   a) Test Length Determination
   b) Item Selection
   c) Layout
   d) Keys
   e) Answer Sheets
8) Determining Standards
9) Test Administration
10) Collection of Reliability, Validity, Norms Data
11) Preparation of User's Manual and Technical Manual
12) Periodic Review of Test

He stressed that it is at step 5a that most commercial publishers fall down. Ron suggested that content people should rate the test items for how well they measure the domains. He also warned that items should not be thrown out because they are too easy or too difficult, since this would bias the item pool.

Ron then addressed the topic of standards. He remarked that all standard-setting methods are arbitrary, but not necessarily capricious. He explained that the paper "Guidelines for Evaluating Criterion-Referenced Tests and

Test Manuals" is a first attempt at putting together a set of guidelines for potential users and for test developers.[8]

Since time was getting short, Ron moved on to describe areas in which he sees a need for further research and development. Well-prepared guidelines for developing domain specifications are not available, he noted, but are needed. Technical standards are also needed. Better test length guidelines would be useful. Ron suggested that latent-trait theory has important implications for CRT. Alternative methods of setting standards should also be investigated, Ron suggested. Such issues as empirical vs. judgemental methods of setting standards, algorithms for comparing methods and procedures to implement the methods might be considered. Work should be done too, he advised, to ensure that classroom teachers are able to use the information reported from criterion-referenced tests. And finally, he noted the necessity of training people to develop tests and instructional materials.

Ron invited people to write to him for more information:

School of Education
University of Massachusetts
Amherst, MA 01003

Ron then entertained questions from the floor. Judah Schwartz stressed that validity should be determined on the basis of children's actual use of the domain skills rather than from putative expertise of subject-matter specialists. Ron responded that collection of information about construct validity occurs at step 10, after the test is constructed, and can include validation against actual use of the skills being measured. Judah Schwartz also expressed his belief that if the analysis of the domain is good enough, and if knowledge of cognitive structures is fundamental enough, then you

end up with a rather different set of performance objectives, and people either can perform or they cannot. This, he explained, leads to bimodal distributions and categorical decisions. Evaluators and educators can then make qualitative rather than what Judah feels are artificial and misleading quantitative decisions. Ron views the structure of domains differently, however. He believes that there are state models which fit Judah's description and which might be pertinent life skills, such as balancing a checkbook. He suggested, however, there are also continuous models, which are still necessary to represent those skills which are best measured by paper and pencil tests. Ron does not feel that academic skills represent an "all or nothing" issue.

Herb Rudman objected to Ron's distinctions between NRT and CRT. Herb maintained that some standardized achievement tests do delineate domains before items are written and that the latent trait model is the approach being taken in at least one major standardized test. Ron countered that differences still exist in the way validity and reliability are assessed for norm- and criterion-referenced tests, since the two testing approaches serve different purposes.

George Hein took exception to Ron's statement that some skills are better measured by paper and pencil tests than by performance tests. George feels that all skills can be evaluated as performance tasks. Since a criticism of NRT has been its restriction to paper and pencil tests, George expressed the hope that CRT would go in the performance direction. Ron replied that the problem is not so much with the test developers as with the consumer who does not want to score the test by hand.

Ann Cook suggested that it was dangerous for advocates of NRT and CRT to assume that they can develop tests based on what consumers want. She urged that the testing technologist ask intelligent questions of consumers and content specialists before developing the tests, questions such as, "Do you really know enough about reading to dissect it into parts which can be tested?"

Peter Airasian asked Ron to comment on two issues: the use of norm-referenced tests in competency-based education, and the method used to set cutoff scores on NRT. Ron responded that, in general, NRT can be used for some criterion-referenced interpretations but generally norm-referenced tests are not so effective as criterion-referenced tests would be for the same purpose. In regard to the setting of cutoff scores, no matter which type of test, the question, "Should the minimally competent person be able to do this item?" should be asked, and this method, he noted, can be applied equally well to NRT and CRT.

Walt Haney observed that the early advocates of CRT proposed more openness in testing and suggested that such tests could be more closely integrated into the teaching and learning process. That attraction does not seem to have materialized, at least in commercially developed tests. Walt asked the reason for this turnaround. Ron replied that, first, most of the commercial tests are objectives-referenced rather than criterion-referenced. Second, Ron continued, he believes that these tests are being used well in the environments in which he has seen them used. He pointed out that they are used by teachers to find out how well they taught certain objectives; they are used as pretests to see which children may be in

special need of help; and they are used by teachers at the end of an instructional segment to determine who ought to go back into remedial work.

IV. Remarks by Banesh Hoffman

Banesh Hoffman voiced concern, on behalf of the Consortium, over the inability of people to write proper English. To illustrate his point, he offered several examples of incorrect grammar which appeared in published material:

> Some teachers quickly read every paper themselves; others read only a portion of them.

Banesh asked, "What does 'them' refer to? What does 'only a portion' mean?" In another instance,

> Some studies were quite elaborate. One involved over 600 high school students who wrote five essays, took six objective tests, and did two other exercises. Each essay was evaluated by 25 readers and two of them by another 145 readers.

Banesh presumed that each student wrote five essays, but he challenged anyone to define "two of them." Another example reads

> Two or three people score every paper and the sum of their scores becomes the paper's final score.

The students whose papers are read by three people, observed Banesh, are lucky.

After citing these examples, Banesh revealed that the irony of these poorly constructed sentences arises from their source: the Educational Testing Service's pamphlet, The Concern for Writing,[9] the purpose of which is to point out how badly students write English. Banesh noted that, after

many years of concentrating on objective test formats, ETS is now pleading for the return of essay tests.

V. Alternative Approaches to Educational Assessment: A Panel

The program for Monday afternoon consisted of a panel on Alternative Approaches to Educational Assessment and an example of an alternative evaluation strategy. The panel was made up of Vito Perrone, George Hein of Lesley College and the North Dakota Study Group on Evaluation, and Fritz Mosher of the Carnegie Corporation.

Vito Perrone introduced the afternoon's topic, alternatives which go beyond norm- and criterion-referenced testing. He noted that concerns about testing have escalated as test use has escalated. People are questioning the reliance on external means of testing.

George Hein explained that he wanted to give the Consortium some examples of alternative assessment strategies being used. First, though, he commented that his framework of assessment strategy comes from a qualitative, holistic, phenomenological perspective. This, he noted, is a different paradigm than the experimental method and NRT and CRT, which arise from an analytic research tradition. George remarked that one can observe behavior, ask about it, or collect products of behavior. He concluded that there is a need for the phenomenological approach to assessment in education. Moreover, this approach must be judged according to the standards of the approach, not according to the standards of NRT or CRT.

George then described the phenomenological approach. It is open-ended, he said, with tests not clearly defined at the beginning and outcomes

frequently different from the objectives. The emphasis of the method is process, the evaluation of performance in appropriate contexts. Multiple methods of data collection are used, and all generally exhibit very high face validity, he asserted.

George drew attention to a number of examples of alternative assessment, most of which are listed on the attached bibliography.[10] George described several of these in more detail to the Consortium.

Brenda Engel (1977) used a variety of assessment methods to examine children's achievement in a third grade in a Cambridge, Massachusetts alternative school, he noted. Carini (1978) developed a very effective system of keeping track of pupils and schools, and she used the assessment products in teacher-training.

George also cited several research-oriented activities using alternative approaches to assessment. The Educational Testing Service has developed a journal, not yet available, titled "Let's Look at Children." The publication is designed to help teachers to assess children's developmental levels. George also directed attention to Goodman's Miscue Analysis, used in assessing reading skills, and work by Churchill Petner of the North Dakota Study Group on Evaluation on children's communications.

In the area of program evaluation, George noted that Duckworth (1978) used materials similar to curriculum materials to assess an African primary science program. Alberty (1977) has developed G.A.M.E., a collection of after-school and in-school activities in which record-keeping strategies assume an active role. Data are collected in a systematic, detailed way, George noted. George also drew attention to Daniel Stufflebeam's qualitative

evaluation of alternative education programs set up in business lobbies, the public library, and on television when Dayton, Ohio was forced to close schools because of the fuel shortage several winters ago. George himself is evaluating programs set up in Boston schools in response to Judge Garrity's order for integration. George explained that schools have been paired with colleges, cultural institutions, and businesses. He noted that a record-keeping system for evaluating the programs has been developed. Both the assessment methods and the results have found acceptance among the programs, the schools, the Massachusetts State Department of Education, and the Massachusetts Board of Education.

Fritz Mosher of the Carnegie Corporation described how Carnegie got into the business of supporting alternatives to standardized testing. He explained that the Corporation's interest was provoked by two streams of thought. First, they became involved with the reanalysis of the Coleman study on the quality of educational opportunity. The results of that study indicated that measured differences in schools were not correlated with differences in measured outcomes. It was reasoned that there must be some schools which make a difference, so there must be some fault in the research. Either the measured differences are not the appropriate ones, Fritz continued, or the measures of outcomes need more attention. The Carnegie Corporation ended up focusing on the latter, Fritz explained. It became clear that the technology used in developing the tests employed in the study was not designed to get at what teachers were likely to teach. It was more probable, according to Fritz, that the standardized tests were measuring correlates of school achievement rather than actual achievement. Fritz observed that test items are

chosen in two ways: according to expert opinion and according to the statistical characteristics of the items. These methods imply that the items are measuring certain things. It seemed to the Carnegie Corporation, however, that to make tests more useful for evaluation, the items should, from the start, be designed directly to measure the attributes of interest. So the Corporation looked into first understanding the skills involved in a subject and then developing tests of the important dimensions of those skills.

The second stream of thought which spurred Carnegie's interest in alternatives was appreciation of the distinction Bob Glaser made between selective and adaptive schooling. In the selective approach, Fritz commented, instruction is presented in a standard form and is either watered down or made more difficult, if necessary. Although children vary in the selective approach to education, the schools remain the same. On the other hand, Fritz explained, the adaptive method varies the nature of instruction for different children in order to achieve commonly high outcomes. Standardized tests, Fritz observed, grew out of the selective approach to education and hence are designed mainly to discriminate among students. In the interest of promoting adaptive education, Carnegie decided to pay more attention to tests which measure the skills being taught and which at the same time record information about the skills in terms of the structure of the subject matter.

Fritz went on to describe four efforts concerning alternative assessment which the Carnegie Corporation is presently supporting. Project Torque, which Judah Schwartz discussed following the panel presentation, is Carnegie-funded. Carnegie also supported development of the New York state reading test which, Fritz acknowledged, fortunately or unfortunately, is going to be

used as a basic competency test of reading for high school graduation. The test itself is interesting for two reasons: first, it takes advantage of the structure of language, and second, it provides grade-equivalent scores which Fritz feels are more meaningful than those currently obtained from standardized tests. For instance, teachers will be able to say that a certain proportion of children in the sixth grade are able to read the materials which are actually used in sixth grade.

A third effort grew out of concern that standardized tests rely too heavily on items requiring students merely to choose among alternative answers, especially in the measurement of writing skills. Fritz noted that the Corporation felt that measures should be of the same genre as the skills. The Corporation came into contact with a group at the High/Scope Educational Research Foundation which had developed a "constructive" language test as part of a Follow Through Planned Variation. Carnegie looked at the test and found it lacking in many respects. Carnegie is now supporting the High/Scope group to conduct a research project dealing with two issues: how to assess the quality of students' written products; and how to judge whether such products are representative of their work.

The New York City accountability program includes plans to set minimum competency standards, to test children on the competencies, and to report the results at the school level. The Carnegie Corporation is supporting the effort there to develop appropriate tests. A combination of paper and pencil tests and actual observation of children reading has been developed, Fritz remarked.

George and Fritz then responded to questions from the floor. Peter

Airasian asked George what kinds of standards are looked for in alternative tests. George replied that there are not any set standards, any more than there are for natural field studies in biology. Rather, a common-sense approach must be used. Peter then asked George to cite the pitfalls of alternative strategies. George quickly noted that, in his opinion, the biggest danger lies in assuming that if the assessment is qualitative, it therefore is easy.

Lillian Weber made the point that another danger is the tendency to collect data over too short a time period. She noted that in order to look properly at a pattern of learning, enough time must be allowed. Ann Cook added that, in order to make sure that interpretations are valid, the qualitative data must be accessible to others.

VI. Alternative Approaches to Educational Assessment: An Example

Judah Schwartz of the Education Development Center and M.I.T. described Project Torque, an example of an alternative assessment strategy. Judah believes that testing has a number of functions, including an instructional function, a role in accountability, and a place in selection and licensure. Project Torque, Judah explained, was developed to fill an instructional need.

Judah listed seven desirable properties of alternative assessment strategies:

1. Assessment should be responsive to individual styles and backgrounds.
2. Assessment instruments should be as free as possible of ethnic bias, and ethnic, cultural, and linguistic assumptions should be clearly stated.
3. The assessment device should be published in its entirety; it should not be secret any more than a curriculum should be secret.

4. An assessment device should be diagnostically useful to both teachers and children.
5. The technology of assessment should be convenient to the child.
6. The assessment situation should be maximally motivating.
7. The assessment procedures should reflect deep understanding of the cognitive structure of the test.

Judah noted three elements of test development which are generalizable beyond a particular subject area. These are the need for understanding of the deep cognitive structure of the skill to be tested, the need for a method of test development in which you work iteratively with the teacher and children, and the need for validation methods which assess the real uses of the skills in question.

Judah explained that he feels there are two types of subjects, the vehicle subjects of math and language, and the tapestry subjects which provide the content in which the vehicle subjects are used. Judah then described in more detail his analysis of the mathematical tasks of measuring and counting. Judah summarized the process of designing a test and validation instruments. First, an analysis of the skill to be tested must be done. This involves watching children using the skills of interest and taking tests designed to measure those skills. Initial tests can then be modified on the basis of such observations, Judah explained. Then you build games which incorporate the skills of interest, such as the measurement of area. Judah stressed that such games must be enjoyable and must enable observation of children actually using the relevant skills. The next step, Judah continued, is to get rid of the developer effect. You give the games to teachers for three weeks, he said, and observe what happens. Then the games can be modified according to

teachers' advice. Finally, he explained, you have a collection of games which become the validating instruments for the paper and pencil tests. Judah stressed, "If you can construct a situation that manifestly engages the child's attention and incorporates in a kind of face valid way the skills you are interested in, that is far, far better to validate than is another paper and pencil test." He added that the paper and pencil test must be validated item by item against the subskill.

Judah pointed out that Project Torque's experience with this process leads to the conclusion that people and skills are more complicated than often assumed. For example, he said, consider the language of time-telling. A child might be able to look at a clock and state correctly that it is "ten to twelve." However, he may write the time as "10:12." The trouble is not in telling time, but in understanding the notation. Judah recounted another instance where a child measured a length of 6.7 centimeters, and wrote it down as "6-7/3." The child explained that there were "six big ones, seven little ones, and three to go." Judah commented that experiences like these are the reason why Project Torque insists on constructed answers rather than multiple choice answers.

Project Torque is developing tests of measurement of length, area, volume, weight, and time. Computation tests have already been developed to cover a five by three grid; encompassing three types of numbers (integers, fractions, and decimals), and five operations (ordering, addition, subtraction, multiplication, and division).[11] Each Torque test measures only one area, for example, addition of integers.

Currently, Project Torque computation materials are being used in pilot

studies with 10,000 children in Washington and New York City. Judah noted that a package of materials on assessment of measurement skills is being finished and will be piloted in the fall of 1979. Project Torque is now beginning work on math-language interaction, that is, word problems. With support from a Ford Foundation grant, Judah also mentioned, a testing advisory service has been set up to help teachers, parents, and school boards think through testing issues.

Discussion followed Judah's presentation. Jean Nazzaro asked Judah to describe the population of children on which the tests were piloted. Judah responded that the group was 1/3 black, 1/3 Latino, and 1/3 Anglo. He commented that this was a deliberate over-representation of minority children but not a preponderance. Deborah Coates expressed concern that ethnicity might be the primary factor in Judah's program. Judah stressed that it was not, but that he did not want Torque to be open to the particular objection of ethnic bias. He went on to explain that one of Project Torque's main concerns is with individual factors. He explained that he had looked at the top three and the bottom three children in a number of classrooms, according to the rankings on standardized tests, and then compared them on the Project Torque tests. He observed that on the Project Torque tests, there may not be so much difference between the children as the standardized tests indicated. For example, if a standardized test over-represents the addition of fractions, then a child may do very poorly in terms of a percentile rank, yet be in the same range of skills in everything except fractions. Or, Judah continued, a child may rank at the 18th percentile on a standardized test because out of 60 items, he completed only 14--but he got them all correct. Because of

such problems, Judah noted, Project Torque tests are not timed.

Peter Airasian asked whether Project Torque was involved in curriculum development or just testing. Judah responded that, in principle, Project Torque does not provide a curriculum, but it points very strongly to one. In fact, people in Israel are using the Torque tests as curriculum materials in 200 Hebrew and Arabic classrooms.

Barbara Lieb-Brilhart asked Judah to describe the implications there might be for teaching the metric system, since that is essentially a new language system. Judah pointed out that since very few children in third, fourth, and fifth grade have strong mental images of inches, feet, and yards, it will be very easy for them to accept the concept of centimeter and meter. Lillian Weber followed up on the issue of metric measurement by noting that the metric system in this country has come from "hard" sciences, yet even in traditionally metric countries people use the vernacular "pinch" rather than "milliliter." There is an historic continuity, she pointed out, between the vernacular and the metric system. Judah concurred that there was a place for both the vernacular and the more precise scientific measures. He remarked that, on the Project Torque "Metric Measuring License," the question is asked of students, "How heavy is the relaxed hand?" Surely, he noted, it is not weightless. Children thus must confront the issue of precision.

In closing Monday's session before the Consortium business meeting, Vito Perrone observed that the sunflower is part of the Project Torque emblem. As part of his ongoing publicity pitch for North Dakota, he informed the meeting that his home state produces 60% of the sunflowers grown in the United States.

VII. Topical Sessions

On Tuesday morning the conference attendants divided into three self-selected groups in order to examine topics of interest more closely and in a more informal setting. Topics for the three sessions were: (1) Alternative Approaches to Educational Assessment, (2) Minimum Competency Testing, and (3) Standards for Tests and Test Use. So that everyone could learn what occurred in the two groups he or she did not attend, the chairmen of the groups summarized the proceedings to the full conference.

Report on Alternative Approaches to Educational Assessment

Vito Perrone, who chaired the group on Alternative Approaches to Educational Assessment, reported that it took a while for the group to get a sense of what was meant in relation to the question of alternatives. The question of purpose in evaluation and assessment was stressed as having to be the base of any discussion. In considering the desirable properties of alternative assessment strategies, Vito noted, the group relied in part on Judah Schwartz's list of criteria presented Monday afternoon. The group quickly focused on one of those criteria, namely that assessment procedures should not be secret. Since a great deal of interest surfaced regarding this issue, Vito suggested that an entire Consortium meeting might be devoted to the topic of accessibility of information regarding testing. Included in such a meeting might be a consideration of what information is available, what is the appropriate use of such information, and what are the philosophical bases for and against test secrecy.

The point was made that alternative approaches to educational assessment should be scrutinized with the same critical eye which now examines traditional

testing methods. It was suggested that a Consortium meeting might be focused on standards in regard to alternative assessment methods.

Vito remarked that an important point brought up in the group session was the recent elementary and secondary education act which calls for state plans to assess oral and written language. The National Council of Teachers of English and the Speech Communication Association stressed that there is no instrumentation which addressed those areas of evaluation. It was suggested, Vito reported, that perhaps the NCT could initiate or support the development of an evaluation strategy in these areas.

Report on Minimum Competency Testing

George Madaus next reported on the Minimum Competency Testing topical session. He outlined a number of issues of concern raised by the group. First, the point was made that two reasons behind the competency testing movement are the accountability movement and the fear that education is not now so good as it was in the past. George observed that the MCT movement has some very positive aspects: it can be used to inform the public better and it can be used as a lever by which the public can change educational practices. However, there is concern over how mandating competency in basic skills areas will affect science, oral communication, and foreign language instruction.

Confusion exists between MCT and competency-based education; these should not be confused, George advised. He continued that the federal government is quite interested in the MCT movement, particularly in three areas. First, it is interested in how MCT is tied to compensatory education. Second, it is concerned with the attitudes of parents and students towards MCT.

Finally, the government is interested in the possible use by lawyers of minimum competency tests as the mechanism for establishing standards of malpractice in education.

Another issue of concern, George noted, is the impact on minority groups, especially in urban areas, of the MCT movement. The movement's effect on curriculum is also of concern. George observed that the MCT movement is seen by some as the deathknell for the inquiry approach to education. The final issue George reported was the concern, "How can you be sure that children who historically have been short-changed by education will be helped and not hurt by the minimum competency movement?"

Several points were raised by the group for the consideration of the full membership, George continued. The group suggested the the NCT push the state of the art in the intelligent use of tests in making broad-gauged assessments about how students are progressing in the basic skills. It was noted that measures other than the judgement of schools and individual teachers are needed and the MCT is one measure that could be a positive help in this regard. The group recommended that the Huron Institute set up a study group to look at the MCT movement on an on-going basis, since it is not sufficient to simply discuss MCT every six months. George advised that consideration of the MCT movement should include its positive as well as negative aspects and also examine what is going on in the classroom in response to the MCT movement.

George concluded with a recommendation emerging from the topical session he summarized that the NCT should study and analyze the possible impact of the Educational Amendments of 1978 under which the federal government will provide

funding for states to establish basic skills assessment programs.

Report on Standards for Tests and Test Use

Walt Haney presented a summary of the discussion in the third topical session, Standards for Tests and Test Use. He described five major areas of group interest and concern. First, he pointed out that the American Personnel and Guidance Association has released a statement on the reponsibilities of users of standardized tests.[12] He continued that, although it would not be appropriate for the NCT to develop its own set of standards, the group did suggest that the Consortium might urge the attention of other groups developing standards toward certain issues such as open availability of tests and the over-reliance on the multiple choice format. The group also recommended that the APA, AERA, and NCME clarify whether the joint standards on testing are intended as indicators of excellence or as minimum requirements of test adequacy. The group also stressed the need for more attention to be focused on issues of test bias and test use with minority groups, the validation of tests with children in "real" settings, and the relation between standards and actual practice.

Second, Walt reported that the Standards group proposed a specific resolution that the NCT try to develop a list of issues important for research to be considered in development of test standards, and that the NCT then instruct the staff to raise these issues with appropriate groups.

Another suggestion introduced in the topical session was that someone undertake a review of certain tests according to a set of standards in order to determine the tests' quality, in a sort of Good Housekeeping Seal of Approval approach. However, Walt remarked that it was felt that such an

effort might dissipate the energies of the NCT to deal with larger issues.

Walt repeated a group member's comment that twenty years more work on test standards might have no impact unless more is done to promote the better use of tests. Walt noted that a resolution would be introduced at the business meeting that afternoon to propose that the NCT undertake the development, publication, and dissemination of a brochure to provide information to legislators and the public on the strengths and weaknesses of tests, in lay terms, and that the content should be widely accepted among professionals.[13]

Finally, the group requested that the staff keep members informed of efforts to revise the joint APA, AERA, NCME Standards and that the staff also provide members with lists of information available from other Consortium members.

VIII. Remarks by Deborah Coates

Deborah Coates, of the Committee on Ability Testing of the National Academy of Sciences, described the purpose of that Committee. It was established, she explained, to carry out a comprehensive review of issues related to ability testing. The committee is gathering evidence regarding the extent of testing, the social context of testing, the present use of testing in major policy decisions, testing's impact on instruction, technical advances in testing, and the impact of recent and proposed legislation on testing. In order to accomplish this review, the Committee is undertaking a variety of tasks, Deborah reported. These include reviewing published and fugitive literature on practices in testing, holding public meetings to gather information from consumers and developers of tests, conducting in-depth interviews

with individuals identified as sources of information on testing practices, and soliciting written testimony on these issues. Deborah noted that the Committee hopes to produce a final report, targeted for the beginning of 1981, which will outline information on these issues and draw policy conclusions and recommendations on what future steps should be taken in regard to testing. Deborah also brought to the Consortium's attention the upcoming public hearing of the Committee on November 17 and 18, to which NCT members are invited to submit written statements.[14]

IX. Report on Huron's Reviews of Major Standardized Test Series

The last topic of the Consortium's substantive meetings was a report on the Huron Institute's reviews of major standardized test series. Walt Haney explained that, under Huron's charge to examine standards for quality of standardized tests, Huron has reviewed the test series most commonly used at the elementary school level. The tests were identified, he recounted, on the basis of a 1978 survey of testing programs in the 100 largest school systems in the U.S., a 1977 report of Title I program evaluations, and a 1975-76 review of tests used in federally sponsored research on children and adolescents.

Walt went on to present ideas for reviewing the quality of these tests. The descriptive review being carried out primarily by Laurie Scott of Huron is simply a starting point, he stressed. Another approach would be to look at the tests in terms of how they are used and interpreted by children in real testing situations, although details for such an analysis have not yet been worked out.

Laurie Scott then outlined the process by which Huron is preparing descriptive reviews of commonly used tests. She noted that Huron first had to assemble the test materials, decide upon a format by which to review the tests, then review the tests, and finally, distribute the reviews. Laurie commented that there have been some problems in obtaining the test materials, although not because of secrecy with regard to the tests. In fact, she noted, only one publisher requested documentation of Huron's qualifications for gaining access to the tests. Instead, problems have been encountered when materials requested from publishers were omitted from a shipment or when publication dates for manuals were delayed.

Laurie then described the outline used for the descriptive review, adapted from outlines suggested by Anne Anastasi and Lee Cronbach.[15] She added that attention has also been given to the joint APA, AERA, and NCME Standards for Educational and Psychological Tests.

Second drafts of all the test reviews have been completed and reviewed internally at Huron, she remarked. Before the reviews are made available to Consortium members, she explained, they will be sent to the tests' publishers for their reactions and comments. Finally, the reviews will form the basis for a review and analysis of commonly used standardized tests, which will be distributed as a Consortium staff circular.

Discussion followed the presentation. Edwin Taylor asked which tests, specifically, were being reviewed. Walt declined to name the tests individually because he preferred to concentrate on the strategy of the reviews at this particular meeting. He explained, however, that the test series were not secret and that Huron would have more substantive information available soon.

Herb Rudman voiced approval of the outline and said he especially liked the fact that Huron was providing the opportunity for feedback from the publishers. He expressed concern, however, that only one person at Huron was doing all the reviews. He asked if any mechanism existed to balance a singular point of view. Walt replied that, although the test reviews were mainly Laurie's responsibility, they were not being done in isolation. Her reviews, he explained, are in turn reviewed by himself and others at Huron, in addition to the publishers.

Roger Farr pointed out that Buros and CSE have already done extensive test reviews. He suggested that instead of repeating these efforts, Huron might be better advised to look at test use and impact in the classroom. Walt replied that the overview will try to analyze the tests in terms of use, but that the descriptive review was merely a starting point for Huron's continuing work in this area. Walt continued that Huron's reviews will in some respects differ from what is available, inasmuch as CSE is strictly evaluative, not descriptive, and Buros' volumes and test catalogs tend to be inconsistent and uneven. Frances Quinto observed that Huron's reviews do draw somewhat on the CSE and Buros evaluations.

Barbara Lieb-Brilhart repeated Roger's sentiments that, since descriptions of the tests are already available, it would be more valuable for Huron to begin by focusing on test use.

Ann Cook countered that, as a practitioner, she would find the reviews quite useful, since Buros and CSE do not address the questions that teachers have. Jean Nazzaro also endorsed the project, especially since the actual testing materials were being studied first-hand.

Gary Kilarr suggested that Huron might ask the test publishers where they saw their tests being used and misused. He observed that the same question could be asked of the 100 largest school districts. Walt expressed interest in this idea and outlined several other review strategies under consideration. These include review in terms of actual test use, validation against the actual use of skills as Judah described the previous day, an analysis of curriculum content implied in the tests, and a systematic consumer's review conducted by talking to school districts' personnel involved in testing programs. Walt continued that there seemed to be a lot of sentiment that the issue of use be a greater focus of Huron's work. A show of hands from the floor indicated this to be the case.

Finally, Frances Quinto requested that at the next conference Huron present a revised timeline and sense of direction regarding the test review plans.

The substantive meetings ended at 12:30 Tuesday afternoon. After lunch, the business meeting begun Monday afternoon was resumed.

Vol. 2, No. 2 December 1976

SLATE

Support for Learning and Teaching of English

MINIMAL COMPETENCIES AND MEASURES OF COMPETENCE

The Issues

Over the past few years there has been an increasing demand at local, state, and national levels for the establishment of a list of essential skills or 'minimal competencies' which every child in school should be expected to attain. The reasons for this demand are many: among them the desire of educational systems to have some sort of accountability system, the dissatisfaction of many with the still large number of citizens who appear unable to read or write, the apparent decline in students' performance on standardized tests, the uneasiness of many with an elective curriculum, and the increasing lack of local support for schools as reflected in tax and bond referenda. As a result certain administrative units have set forth requirements for verified competencies of secondary school students—states like Arizona, California, Oregon and communities like Gary (Indiana), Duval County (Florida), Westside Schools (Omaha, Nebraska) are but a sample of the many states and districts using minimal competency tests. At the national level, the National Association of Secondary School Principals has endorsed the recommendation of its Task Force that there be uniform graduation requirements verified by competency measures as well as by units or credits.

The implications of this movement are not entirely clear, but certainly it raises a host of potential problems. These include:

1. the possibility that scores on competency tests will determine promotion or non-promotion and thus will lead to a return to grade-repeating, a practice which disappeared from American schools after research pointed to its ill-effects on students and on schools;
2. the possibility that there will emerge diplomas of different 'classes' dependent on the student's scores on a test, with attendant danger of social stigmatism;
3. the possibility that statements of competence will lead to a circumscribing of the curriculum to a point where it will be merely preparing students for the test rather than educating them broadly;
4. the possibility that statements of competence will lead to didactic teaching as a shortcut to improving student performance, when research has indicated that such teaching is not as beneficial as more varied approaches to teaching;
5. the possibility that educational resources will be concentrated upon those students who have trouble attaining the minimal level of competence, and will be diverted from the gifted and the broad range of students who can attain the minimal level fairly easily but need to go beyond it;
6. the possibility that educators and the public will be satisfied with bringing students to a minimal level of competency rather than to a level of mastery that is appropriate to their age, their view of themselves, and their aspirations;
7. the possibility that statements of competence and measures of competency will make schools less responsive to the cultural and linguistic diversity of this country.

Obviously these implications bear social and political dimensions which can only be explored through observation of the long-term effects of competency programs and tests.

Professional Viewpoints: NCTE/Research

There has been a great deal of study of various aspects of testing as well as of aspects of goal-setting, competency definition, and behavioral objectives.

The research on testing is too extensive to be dealt with justly in a brief space; its major areas have concerned reliability (how effective a test is in measuring a phenomenon accurately) and validity (how well a test measures a phenomenon or acts as a predictor of subsequent behavior). Reliability studies usually deal with whether a test is internally consistent, whether it measures a student's performance the same way when it is given to the student again, or whether two versions of the same test produce the same scores when a particular student takes them. Reliability research indicates that standardized tests are generally consistent in one of these ways, but such research does not indicate anything about the usefulness of the tests concerned.

Validity studies seek to determine whether an objective test in writing, for example, is as good a measure of writing performance as is the grading of a set of papers from a particular student. Some validity studies use experts to examine the test and to say whether the questions are dealing with appropriate material or following some established theory; such studies determine the "content validity" of the test. Examples of such studies can be found in O. Buros' *Mental Measurements Yearbook* (N.J.: Gryphon Press) and A. Grommon, *Review of Selected Published Tests in English* (Urbana: NCTE, 1976). The other form of validity research is called "empirical" or "criterion" validity. It seeks to determine whether a test measures some attribute of learning that can be observed in practice. A test of reading interests might be validated against the library withdrawals of a particular

student, or a test of spelling against a count of spelling errors in a student's writing. Validity studies are much more difficult to construct than are reliability studies; very few get beyond content validity to empirical validity. A major validity study of the College Entrance Examination Board's Achievement Test in English showed that the test was somewhat valid but that its validity could be increased by the addition of a piece of writing that was graded by trained readers (Goldschalk, Swineford, and Coffman, *The Measurement of Writing Ability*. New York: CEEB, 1966). Other empirical validity studies can be found in various articles in *Research in the Teaching of English*, and in Fagan, Cooper, and Jensen, *Measures for Research and Evaluation in the English Language Arts* (Urbana: NCTE and ERIC/RCS, 1975).

As far as can be determined, there have been no empirical or criterion validity studies of competency tests.

Research in competency statements is generally meagre, though they could be validated against the opinion of experts or against the kinds of tasks actually performed by competent citizens. In mathematics, such a validation might begin with a look at the kinds of mathematics people employ in different work and family situations, and continue with an analysis of what might be general levels of competence and what might be more specialized levels (for a mechanic, say, or an engineer). In English, a similar validation procedure could take place. What kinds of writing do people actually do? Although competence statements have been attempted for the National Assessment of Educational Progress and in the course of studies of mastery learning (J.H. Block, *Mastery Learning* [N.Y.: Holt, Rinehart, and Winston, 1971]), they have come from experts' opinions rather than from a careful scrutiny of what goes on in the world.

Strategies for Action

Teachers in those states or districts that are considering legislation related to competency programs have a different problem from those in locations where legislation already exists. In the first case, teachers and teachers' groups have an opportunity to lobby concerning the direction of legislation. A useful source for potential lobbyists is the publication of the American Library Association, *Who Me, A Lobbyist?*, available from the ALA Washington Office, 110 Maryland Avenue N.E., Washington, D.C. 20002. In the second case, teachers have to work to make sure that the criteria of competence and the measures of competence meet their own professional standards. The following guidelines, drafted by an NCTE *ad hoc* Committee on Minimal Competency Testing, may provide a useful starting point.

GUIDELINES FOR THE DEVELOPMENT OF COMPETENCY STATEMENTS AND MEASURES OF COMPETENCE IN ENGLISH

Competency Statements Must Be Sufficiently Comprehensive So as to Cover the Many Facets of English

1. Competency statements in English must include the broad areas of reading, writing, oral language, and media literacy.

2. Competency statements in English must include a variety of contexts (e.g., statements should not be limited to business communication).

3. Competency statements should emphasize those aspects of English which are critical in effective communication, as opposed to those which might be trivial or less significant (e.g., clarity of expression in a letter of application is more critical than using the appropriate state abbreviation in an address).

4. Competency statements in English must reflect the developmental level of the students (e.g., criteria for making inferences about reading must consider the potential ability of students at a given age to make inferences of a particular level of abstraction).

Measures of Competence in English Must Be Many and Varied

1. Measures of student competence in English must include not only tasks dealing with analysis but tasks calling for performance in oral and written language.

2. Measures of student competence in English must include tasks that are to be evaluated for their overall effectiveness by teachers or other professionals in English.

3. Measures of student competence in English should include language presented orally, in print, and through electronic media.

4. Measures of student competence in English must include tasks which are to be observed and evaluated by English teachers on a continuing basis (e.g., there should be some report on the student's use of language in daily situations as well as in specific test situations).

5. Measures of student competence in English must include provisions for superior performance (e.g., a measure of student performance in writing must include the possibility that a student could be judged an excellent writer so that incentives to excellence can remain in the curriculum and the testing process).

Alan C. Purves
(For the SLATE Steering Committee)

National Council of Teachers of English
1111 Kenyon Road
Urbana, Illinois 61801
(217) 328-3870

The SLATE STEERING COMMITTEE offers the Starter Sheets as resources for dealing with current issues affecting the teaching of English. Reproduce these sheets and use them in any ways that might help to promote better understanding of the goals of English teaching.

X. FOOTNOTES

[1] A summary of this bill is appended to this paper.

[2] These strategies are described in more detail in the attachment, "The Testing, Assessment and Evaluation Program," and so will not be repeated in this summary.

[3] These strategies are described in detail in Jeff Schiller's paper attached.

[4] This is described in the attached paper, "A Brief Description of the American Council on Education Credit-by-Examination Test Review Program," which formed the basis for Doug Whitney's presentation.

[5] The December 1976 SLATE bulletin on Minimum Competencies and Measures of Competence is attached to this summary.

[6] More information on the court challenge is available in Education Daily, Vol. 11, No. 202, October 18, 1978.

[7] Ron Hambleton presented an extensive bibliography on CRT to the meeting. The articles which Ron specifically recommended are listed in the attached bibliography. Additional references are available from Ron.

[8] A copy of this paper is attached.

[9] The Concern for Writing, ETS FOCUS 5, can be ordered by writing to: FOCUS; 112 Thurstone Hall; Educational Testing Service; Princeton, N.J. 08541.

[10] Those references which George mentioned specifically in his talk are marked with an asterisk.

[11] Samples of these tests and a "Briefing Document on Torque Mathematics Assessment Materials" are available from: Judah Schwartz; Education Development Center; 55 Chapel Street; Newton, MA 02160.

[12] A copy of the APGA statement is attached.

[13] This resolution was discussed at the business meeting Tuesday afternoon. The Consortium decided to direct the Huron Institute to develop a specific proposal regarding the scope and content of such a brochure. Huron's proposal will then be presented to the Consortium membership.

FOOTNOTES continued...

[14] More detailed information on the hearing is provided on the attachment and so will not be repeated here. Also attached are a description of the Committee and a list of Committee members and staff.

[15] A copy of the outline is attached to this summary.

Senate Bill No. 2005

CHAPTER 1217

An act to add Chapter 3 (commencing with Section 99150) to Part 65 of the Education Code, and to amend Section 6254 of the Government Code, relating to postsecondary education.

[Approved by Governor September 26, 1978. Filed with Secretary of State September 26, 1978.]

LEGISLATIVE COUNSEL'S DIGEST

SB 2005, Dunlap. Postsecondary education: standardized tests.

Existing law does not regulate the activities of testing services which prepare and administer standardized tests for various purposes related to education.

This bill would require such testing services, defined generally as test sponsors, to make various reports to the California Postsecondary Education Commission regarding the administration and use of standardized tests to California inhabitants. Such reports would relate to both the operation and charges for the services. Such material would be subject to disclosure under the California Public Records Act. Test sponsors would also be required to file copies of specified test questions and answers with the commission.

This bill would require test sponsors to file certain information sufficient to describe the psychometric quality of the test.

This bill would require test sponsors to notify California test subjects regarding the purposes of the test, the subject matter of the test, the manner in which the test is scored, the manner in which the test scores will be recorded, the basis upon which scores will be made available, and a sample of test items.

This bill would make any test sponsor who intentionally violates the bill liable for a civil penalty of not to exceed $750 for each violation.

This bill would incorporate additional changes in Section 6254 of the Government Code proposed by Assembly Bill No. 1091, to be operative on the operative date of Assembly Bill No. 1091 only if Assembly Bill No. 1091 and this bill are both chaptered and become effective on January 1, 1979, and this bill is chaptered last.

The people of the State of California do enact as follows:

SECTION 1. The Legislature finds and declares that:

(a) Because standardized tests often play a major role in the admission, placement, and evaluation of students at all levels of our educational system, students, parents, and the general public have the right to know and to understand the nature, purposes, limitations, uses and potentials of all kinds and types of standardized

DRAFT 1/

-- Jeff Schiller --

The Testing, Assessment and Evaluation Program

I. Introduction

The Testing, Assessment and Evaluation Program is concerned with assessing and improving educational achievement and competencies by improving the value of both testing and evaluation. While the potential educational value of both testing and evaluation in the schools is high, neither serves yet as an effective tool for improving the quality of education; nor is there consensus on the role of each.

In the testing area, the use of test scores for school entry, grade placement, school exit and occupational selection is vigorously criticized by groups such as professional teachers organizations and concerned parent groups. Norm-referenced tests in particular are criticized as irrelevant to instructional decisions. At the same time, employers complain that high school graduates do not have basic skills necessary for holding jobs and colleges complain about the need for remedial courses. One widespread result is that forces outside the schools, including school boards and state legislatures, are implementing minimum competency testing programs as a means of ensuring that students meet minimal standards of performance

---

1/ This paper represents an initial draft of a framework for guiding the development of the TAE program. All comments and suggestions for improvement are welcomed.

in certain subjects before graduating from high school.

The role and purpose of program evaluation has also been questioned. While administrators do need to know which of their programs or strategies are successful under given conditions, very few evaluations have yielded information that has proved useful in informing their decision-making needs. As financial resources become more scarce, the need for such evaluative information continues and intensifies.

Within this context the program on Testing, Assessment and Evaluation sponsors research, development and dissemination activities which address the following two objectives:

For Testing: To identify more effective means of assessing and improving student performance through improved testing;

For Evaluation: To evaluate and actively disseminate information about promising educational programs.

II. Testing

To accomplish its objective, the Testing Program will pursue three strategies: improving current testing practices, clarifying the role and impact of testing in education, and promoting the integration of testing with instruction and learning.

Examples of activities for each strategy are described below:

Strategy 1. Improving testing practices:

-- developing and disseminating easily understood materials on testing for parents and teachers;

-- sponsoring regional workshops and conferences on critical issues in testing;

-- identifying and disseminating models of exemplary testing programs and prototypes for test development;

-- exploring ways of reducing the testing burden in the schools;

-- exploring the use of test scores and other performance measures as "indicators" of educational progress;

-- investigating problems in test development, test validity, analysis of test scores, and interpretation of test results, especially in relation to minority group concerns;

-- exploring the effects of cultural and linguistic demands of tests, and proposing ways of estimating and reducing the degree of cultural and linguistic demands of tests.

Strategy 2. Clarifying the role and impact of testing in education:

-- exploring the uses and misuses of testing;

-- studies of the social context of tests and testing;

-- evaluating the impact of State and Local testing programs, including minimum competency testing programs.

Strategy 3. Integrating testing with instruction and learning:

-- examining how tests can be constructed to identify what students know and don't know, to analyze patterns of errors, and to identify other process information which will help the student learn and the teacher teach;

-- proposing new assessment strategies for integrating testing information with instructional decisions;

-- investigating alternative forms of testing that capitalize on recent advances in computer technology and research in cognitive processes and learning theory;

-- examining the instructional value and psychometric properties of content-referenced tests.

III. Evaluation

To accomplish its objective, the Evaluation Program will pursue three strategies: identifying and evaluating promising programs that can improve the literacy skills and competencies of target groups, advancing the state of the art of evaluation methodology,

and strengthening the capacity of Federal, state and Local organizations to conduct evaluations of their own programs. Examples of activities for each strategy are described below:

Strategy 1: Evaluation of promising programs:

-- evaluating a small number of selected educational innovations with potential national significance, e.g., Cities in Schools and EXCEL;[2/]

-- evaluating a variety of educational programs in existence in cities designed to cope with the blight and challenges to learning posed in urban education;

-- through integration of findings from the two activities just mentioned plus those from other appropriate evaluations, constructing "meta" or broadly-based syntheses of effective practice in urban education;

-- exploiting large scale evaluative data bases already in existence (especially longitudinal ones) capable of yielding new information about effective educational practice, through secondary analysis.

2/ While these evaluations may be initiated by the staff, in many cases they will be in response to initiatives originating outside NIE. Such initiatives might typically come from departmental levels in HEW. In this capacity, the evaluation program is serving larger interests in educational innovation, including those in which more than one Federal agency may have an interest.

Strategy 2: Advancing the state of the art of evaluation methodology:

-- addressing pressing technical issues, e.g., equating noncomparable treatment and control groups and alternatives to quasi-experimental designs in field work, including ways of using ethnographic and other qualitative data for analyses and decisionmaking;

-- exploring ways of improving the use of evaluation findings by decision-makers through increased responsiveness of evaluation design, involvement of decision-makers and target groups in the evaluation process, and more effective communition and explanation of findings;

Strategy 3: Strengthening the capacity of Federal, state and local organizations to conduct evaluations:

-- generating information about the extent and functioning of evaluation offices in local school districts and the practice of program evaluation at the local level;

-- providing timely information regarding new developments in evaluation to state and local education agencies;

-- developing means to reduce evaluation requirements engendered by numerous Federal and state programs;

-- developing and disseminating professional standards for educational evaluation.

IV. Comments

The research described above will reflect various methodologies, disciplines, and approaches. The projects will consist of a mix of activities ranging from the creation of new knowledge to applied research to synthesis and dissemination. While the mix of activities, and priorities may vary over time, the six research strategies described above will remain basically the same.

Table 1 displays an overview of TAE in a shorthand form and provides a guide for current projects under each strategy. TAE will continue to review and modify the mix of projects both within and among the research strategies to reflect current priorities and concerns.

Table 1

The Testing, Assessment, and Evaluation Program

FY 79

Goals of The Teaching and Learning Group	The Testing, Assessment and Evaluation Division	Division's Overall Objectives	Division Strategies for Achieving Its Objectives	On-going Projects Responsive to each strategy (numbers refer to numbered abstracts in "Index of Projects"
To improve literacy across all age groups and in diverse settings →	TESTING →	Identifying more effective means for assessing and improving student performance	1. Improving current testing practices 2. Clarifying the role and impact of testing in education 3. Integrating testing with instruction and learning	100 - 116 200 - 205 300 - 303
	EVALUATION →	Evaluating and actively disseminating information about those educational programs that can be exemplary in terms of achieving their stated objectives.	1. Identifying successful programs through field evaluations and secondary analyses of data 2. Advancing the state of the art in evaluation methodology 3. Strengthening the ability of Federal, State, and Local organizations to conduct evaluations of their own programs	400 - 409 500 - 524 600 - 602

--Doug Whitney--

A BRIEF DESCRIPTION OF THE AMERICAN COUNCIL ON EDUCATION CREDIT-BY-EXAMINATION TEST REVIEW PROGRAM

Basis for presentation at the National Consortium on Testing Fall Conference, Washington, D. C., October 23, 1978

Background

The American Council on Education has, for many years, been actively involved in the development and use of methods for assessing the knowledge, skills and competencies of students entering and enrolled in this country's postsecondary education institutions. The Council is now, as it has been throughout this phase of its work, committed to two principles. First, the assessment procedures used for awarding credit and for satisfying credentialing requirements should be of the highest quality and should be used in a responsible manner. Second, persons who have acquired significant levels of relevant knowledge, skills, and competencies outside of regular institutional educational programs should have ample opportunity to demonstrate their achievements through means that are appropriate bases for awarding college-level credit and satisfying credentialing requirements.

Probably the most widely-known Council activity in this area is the General Educational Development (GED) testing program. This program, administered since 1945 by the Council's Veteran's Testing Service and since 1963 by our GED Testing Service, has enabled more than a million persons to earn high school equivalency credentials; nearly 350,000 did so in 1977 alone. Our role in the GED program is twofold: ACE develops high quality examinations and nationally representative norms for the GED; ACE supervises the administration of the tests at 2,704 official GED centers. Each state, six territories or possessions, and most Canadian provinces have established their own requirements for equivalency certification based on GED test results.

For a number of years, the Council also operated a college-level GED testing program. This program was succeeded by the Educational Testing Service's Comprehensive Testing Program in 1961 and later by the College Entrance Examination

Board's College Level Examination Program (CLEP). The Council was also instrumental in establishing the Educational Testing Service.

For a number of years, the Council's Commission on Educational Credit has made recommendations to postsecondary institutions about the kind and amount of college-level credit that might be appropriate for examinees earning high scores on the CLEP examinations and on the United States Armed Forces Institute (USAFI) examinations (now administered by the Defense Activity for Non-Traditional Education Support -- DANTES). Many postsecondary education institutions base their credit-by-examination policies on these recommendations.

This brief history of the Council's involvement in testing activities at the post-secondary level is offered to inform those of you who might not be familiar with this phase of the Council's activities. It is also intended to provide a context in which the Credit-By-Examination Test Review Program can be recognized as a logical out-growth of earlier efforts. That is, the review program I am about to describe does not represent a wholly new program of a "Johnny-or-Janie-come-lately" organization, but an extension and refinement of Council activities dating back more than thirty years.

Current Council Concerns

The Council's Review Program was developed in response to four concerns about the policies of postsecondary institutions in the credit-by-examination area. In no particular order, these are:

(1) The current Council recommendations for the CLEP and USAFI examinations are in danger of becoming outdated due to extensive changes in the examinations and to the lengthy period that has elapsed since they were originally issued.

(2) The Council believes it desirable to establish an ongoing basis for such recommendations in order to insure their currency and to extend them to other examinations and assessment procedures.

(3) It has been suggested that some institutions have adopted and/or set standards for external examinations without the substantial involvement of local faculty -- an involvement that the Council considers essential to the legitimacy of credit-by-examination programs.

(4) It is clear that some institutions have established standards for granting credit on such examinations solely on the basis of Council recommendations, without the kind of detailed local review that the Council strongly encourages and/or without a clear rationale and equitable basis for the standards that are established.

The Test Review Program was established in order to address these concerns and to better enable the Council to:

(1) improve the quality of examinations and institutional policies used in credit-by-examination programs, and to

(2) expand the opportunities for persons who have acquired substantial levels of knowledge, skills, and competencies in extrainstitutional settings to receive appropriate credit and credentialing recognition.

While, at first glance, these goals may seem antagonistic, the Council believes that they are not. In fact, we believe that the two can be mutually supportive if addressed together. That is what the Credit-By-Examination Test Review Program attempts to do.

Outline of the Test Review Program

The program involves a number of activities that are directed toward helping postsecondary education institutions develop high quality credit-by-examination programs. The Office on Educational Credit will publish a Guide to Credit-By-Examination listing and describing tests recommended for use in postsecondary institutional credit-by-examination programs. Each listing will represent a test (or test program) voluntarily submitted by the test publisher and reviewed favorably by teams of subject matter and technical specialists selected by ACE for that purpose. The listings will offer recommendations about the

type and amount of college-level credit appropriate for examinees achieving a specified minimum credit awarding score on each test. The publication will be revised and reissued periodically as additional reviews are completed and as past recommendations expire. A companion publication, Credit-By-Examination Programs: Recommendations for Reviewing Tests and Setting Standards, will discuss issues to be resolved and will offer specific advice on procedures to be used in establishing comprehensive credit-by-examination policies. Examples will be drawn from the guidelines and procedures used in the ACE test reviews as well as from exemplary institutional efforts. OEC staff will assist local institutions in developing regional and local resources for conducting test reviews and for implementing these recommendations.

Selected Details of the Reviews

Permit me to highlight a few specific elements of the review process that will, I hope, offer some insight into the care that has been exercised in developing this Program and suggest the high standards that we feel will be reflected in the resulting activities.

This review process is to be an ongoing activity. Although the initial Guide to Credit-By-Examination publication is scheduled for release in fall 1979, periodic updates or revisions will be issued. Each approved test listing will include a specific expiration date for the Council's recommendation. If, by the time the Guide is next revised, an expiration date has passed without a new review having been completed, the test will be dropped from the listing.

The review team charged with evaluating the content of each test will be composed of three experienced faculty members who have taught at accredited institutions in the subject area(s) covered by the test and who were nominated for this task by officers of relevant national or regional professional associations. They will evaluate the content validity, the currency, representativeness, and clarity of the examination

tasks or items. In addition, they will offer recommendations about the amount and level of credit that seems appropriate, the minimum performance level or score that should be required in order to award credit, and the period for which the approval (if recommended) should be granted. A specific discussion of these elements is contained in the Guidelines for Reviewing the Content of Examinations for Granting College Credit available from the Council's Office on Educational Credit.

The teams that evaluate the technical quality of each examination are composed of three persons with considerable educational or experiential training in measurement, psychometrics or related areas who are also experienced in test development. These persons are selected from the nominees of the officers of relevant professional organizations (e.g., National Council on Measurement in Education, American Educational Research Association). A discussion of the specific and general considerations covered in these reviews is available from the Council's Office on Educational Credit in the Guidelines for Reviewing the Technical Quality of Examinations for Granting College Credit.

The Guide will also contain a Council recommendation on the minimum level of achievement that should be required before awarding credit based on examination or other assessment methods. Although the wording has not yet been finalized, we expect the policy to emphasize the desirability of ensuring that the level of achievement required for awarding course or credentialing credit by these means is comparable to that required of students satisfactorily completing institutional equivalents. When possible, one or more specific score recommendations consistent with this policy will be listed in the Guide for each test approved.

Douglas R. Whitney
Associate Director,
Office on Educational Credit

NCME
7/21/78

Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals

Ronald K. Hambleton and Daniel Eignor
University of Massachusetts, Amherst

Abstract

The scope and number of criterion-referenced tests available to potential users is impressive. Unfortunately, the quality of these tests varies tremendously and so it is very important for potential users to carefully review available tests before making their selections.

The primary purpose of this paper is to propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests . A secondary purpose is to report on our use of the guidelines with eleven commercially available criterion-referenced test batteries.

Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals[1,2,3]

Ronald K. Hambleton and Daniel R. Eignor
University of Massachusetts, Amherst

Most of the major test publishers have published in the last few years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced tests. Criterion-referenced tests are designed to address many problem areas. For example, criterion-referenced tests are being used to monitor student progress through school programs, to diagnose learning disabilities, to report student progress to parents, to evaluate various types of programs, and to certify or license professionals in many fields. Unfortunately, it appears to us, and to many users of criterion-referenced tests we have spoken with, that many of the available tests fall short of the technical quality necessary for them to accomplish their intended purposes. Perhaps one explanation is that many criterion-referenced tests were developed before an adequate testing technology was fully explicated. Fortunately, there now exists an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores (Hambleton and Eignor, 1978; Hambleton, Swaminathan, Algina, Coulson, 1978; Popham, 1978). Another possible explanation is that there has been a shortage of guidelines for constructing and using criterion-referenced tests. Certainly the well-known Test Standards for

---

[1]Paper presented at the annual meeting of NCME, Toronto, 1978.

[2]Laboratory of Psychometric and Evaluative Research Report No. 73. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1978.

[3]A shorter version of this paper will appear in the Journal of Educational Measurement, 1978, in press.

evaluating tests and test manuals prepared by a joint committee of AERA/APA/NCME is helpful, but it is not completely applicable to criterion-referenced tests. Besides the incompleteness of the AERA/APA/NCME Test Standards for evaluating criterion-referenced tests and test manuals, what relevant information there is, is scattered through 75 pages or so of other materials appropriate for norm-referenced test evaluations. Therefore, the Test Standards in its present form, is not very useful for individuals interested in evaluating criterion-referenced tests.

The primary purpose of this paper is to propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests. Test standards are not offered in the paper (an example of a standard is, "test score reliability must exceed .80"), but we do offer a set of questions for consideration by potential users and developers of criterion-referenced tests. The only other efforts we are aware of to develop guidelines for evaluating criterion-referenced tests and test manuals are Popham (1978, Chapter 8), Swezey and Pearlstein (1975), and Walker (1977). A secondary purpose is to report on our use of the guidelines with eleven commercially available criterion-referenced test batteries.

One caution and one comment seem appropriate to introduce at this point. The guidelines represent our own biases about what is important technical information for users to have in making informed decisions about the quality of criterion-referenced tests. Also, in this paper we did not provide (1) a rationale for the inclusion of each guideline, and (2) specifics on how the guidelines were applied. Interested readers

are encouraged to read Eignor (1978) and Hambleton and Eignor (1978) for the information.

A Proposed Set of Guidelines

The list of guidelines was generated by placing ourselves in the role of potential purchasers of a criterion-referenced test, and asking "What questions would we want to answer before making a decision to use a criterion-referenced test in a particular situation?" Questions were organized around ten broad categories[1]. They are: Objectives, Test Items, Administration, Test Layout, Reliability, Cut-off Scores, Validity, Norms, Reporting of Test Score Information, and Test Score Interpretations. The questions are as follows:

Objectives

A.1 Is the purpose (or purposes) of the test stated in a clear and concise fashion?

A.2 Is each objective clearly written so that it is possible to identify an "item pool"?

A.3 Is it clear from the list of objectives what the test measures?

A.4 Is an appropriate rationale offered for including each objective in the test?

A.5 Can a potential user "tailor" the test to meet local needs by determining which objectives from a pool of objectives offered by the publisher are to be measured by the test?

A.6 Is there a match between the content measured by the test and the situation where the test is to be used?

A.7 Are individuals identified who were responsible for the preparation of objectives?

A.8 Does the set of objectives measured by the test serve as a representative set from some content domain of interest?

---

[1]The very important factors of cost and time limits are not considered here, but they are included in our evaluation form.

B. Test Items

B.1 Is the item review process described?

B.2 Are the test items valid indicators of the objectives they were developed to measure?

B.3 Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?

B.4 Are the items free of technical flaws?

B.5 Are the test items in an appropriate format to measure the objectives they were developed to measure?

B.6 Are the test items free of bias (for example, sex, ethnic, or racial)?

B.7 Was a heterogeneous sample of examinees employed in piloting the test items?

B.8 Was the item analysis data used only to detect "flawed" items?

C. Administration

C.1 Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?

C.2 Are the test directions clear?

C.3 Is the test easy to score?

C.4 Does the test manual specify an examiner's role and responsibilities?

D. Test Layout

D.1 Is the layout of the test booklets attractive?

D.2 Is the layout of the test booklets convenient for examinees?

E. Reliability

E.1 Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?

E.2 Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?

E.3 Are test lengths suitable to produce tests with desirable levels of test score reliability?

E.4 Is reliability information offered in the test manual for each intended use (or uses) of the test scores?

F. Cut-Off Scores

F.1 Was a rationale offered for the selection of a method for determining cut-off scores?

F.2 Was the procedure for implementing the method explained, and was it appropriate?

F.3 Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?

G. Validity

G.1 Does the validity evidence offered in the test manual address adequately the intended use (or uses) of scores obtained from the test?

G.2 Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?

H. Norms

H.1 Are the norms data reported in an appropriate form?

H.2 Are the samples of examinees utilized in the norming study described?

H.3 Are appropriate cautions introduced for proper test score interpretations?

I. Reporting of Test Score Information

I.1 Are the test scores reported for examinees on an objective by objective basis?

I.2 Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?

I.3 Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?

J. Test Score Interpretations

J.1 Are suitable cautions included in the manual for interpreting individual and group objective score information?

J.2 Are appropriate guidelines offered in the manual for utilizing test scores to make descriptive statements, instructional decisions, program evaluation decisions, or other stated uses of the test scores?

A convenient rating form is given on the next four pages.

Evaluation of Eleven Criterion-Referenced Tests

Eleven of the more popular criterion-referenced tests were selected for review. The names of the tests and some descriptive information are presented in the chart.

INSERT THE CHART ABOUT HERE.

Our primary purpose was to ascertain the extent to which these tests met our guidelines. We have reported our evaluation of each test relative to each guideline, but the more important information is arrived at by determining how well the tests as a group meet each of our guidelines. The group information is informative because it helps to pin-point areas where commercial materials are in need of revisions and further development.

Criterion-Referenced Test and Test Manual Evaluation Form

Background Information

Test Name:_____ Forms and Levels:_____

Test Publisher:_____ Author(s):_____

Year of Publication:_____ Cost:_____

Reusable Booklets: Yes No

Special Test Administration Conditions:_____

Manual and Other Technical Aids:_____

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question.

Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.1. Is the purpose (or purposes) of the test stated in a clear and concise fashion?					
A.2. Is each objective clearly written so that it is possible to identify an "item pool"?					
A.3. Is it clear from the list of objectives what the test measures?					
A.4. Is an appropriate rationale offered for including each objective in the test?					
A.5. Can a user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?					
A.6. Is there a match between the content measured by the test and the situation where the test is to be used?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.7. Are individuals identified who were responsible for the preparation of objectives?					
A.8. Does the set of objectives measured by the test serve as a representative set from some content domain of interest?					
B.1. Is the item review process described?					
B.2. Are the test items valid indicators of the objectives they were developed to measure?					
B.3. Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?					
B.4. Are the items free of technical flaws?					
B.5. Are the test items in an appropriate format to measure the objectives they were developed to measure?					
B.6. Are the test items free of bias (for example, sex, ethnic, or racial)?					
B.7. Was a heterogeneous sample of examinees employed in piloting the test items?					
B.8. Was the item analysis data used only to detect "flawed" items?					
C.1. Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
C.2. Are the test directions clear?					
C.3. Is the test easy to score?					
C.4. Does the test manual specify an examiner's role and responsibilities?					
D.1. Is the layout of the test booklets attractive?					
D.2. Is the layout of the test booklets convenient for examinees?					
E.1. Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?					
E.2. Was the sample of examinees adequate in size, and representative of the population for whom the test is intended?					
E.3. Are test lengths suitable to produce tests with desirable levels of test score reliability?					
E.4. Is reliability information offered in the test manual for each intended use (or uses) of the test scores?					
F.1. Was a rationale offered for the selection of a method for determining cut-off scores?					
F.2. Was the procedure for implementing the method explained, and was it appropriate?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
F.3. Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?					
G.1. Does the validity evidence offered in the test manual address adequately the intended use (or uses of scores) obtained from the test?					
G.2. Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?					
H.1. Are the norms data reported in an appropriate form?					
H.2. Are the samples of examinees utilized in the norming study described?					
H.3. Are appropriate cautions introduced for proper test score interpretations?					
I.1. Are the test scores reported for examinees on an objective by objective basis?					
I.2. Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?					
I.3. Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?					
J.1. Are suitable cautions included in the manual for interpreting individual and group objective score information?					
J.2. Are appropriate guidelines offered for utilizing test scores to accomplish stated purposes?					

Tests Selected for Review

Code	Name of Test	Grades	Levels	Forms	Publication Date	Publisher
1	1976 Stanford Diagnostic Mathematics Test	1-12	4	2	1976	Harcourt Brace Jovanovich
2	1976 Stanford Diagnostic Reading Test	1-12	4	2	1976	Harcourt Brace Jovanovich
3	Skills Monitoring System-Reading	3-5	3	1	1975	Harcourt Brace Jovanovich
4	Individual Pupil Monitoring System-Mathematics	1-6	6	2	1974	Houghton-Mifflin
5	Individual Pupil Monitoring System-Reading	1-8	8	2	1974	Houghton-Mifflin
6	Diagnostic Mathematics Inventory	1.5-7.5	7	1	1977	CTB/McGraw-Hill
7	Prescriptive Reading Inventory	K-6.5	6	1	1977	CTB/McGraw-Hill
8	Diagnosis: An Instructional Aid-Mathematics and Reading	1-6	2	2	1974	Science Research Associates
9	Mastery: An Evaluation Tool-SOBAR Reading	K-9	10	2	1975	Science Research Associates
10	Mastery: An Evaluation Tool-Mathematics	K-8	9	2	1974	Science Research Associates
11	Fountain Valley Support System in Mathematics	K-8	9	1	1974	Richard L. Zweig Associates

In judging the quality of a test and test manual relative to each guideline, the following rating scale was used:

A	= Acceptable
A⁻	= Acceptable, with reservations
X	= Unacceptable, data offered was unsuitable or improperly used
Y	= Unacceptable, no data was offered
N	= Not Applicable

Table 1 summarizes our ratings of the 11 tests on the 39 guidelines.

INSERT THE TABLE ABOUT HERE

Our most significant impressions of the test and test manuals reviewed are as follows:

1. In areas such as Administration, Test Layout, and Norms, there are few problems.

2. Current commercially available "criterion-referenced tests" reviewed in this paper should be called "objectives-referenced tests" since the tests appear to be developed from behavioral objectives (Popham, 1978). Starting to develop a test from a listing of behavioral objectives is less than ideal because behavioral objectives usually do not lead to unambiguous definitions of the "item pools" keyed to the behavioral objectives. The solution is to write "domain specifications" (Popham, 1978).

3. Only about half of the publishers included information about the qualifications of individuals who prepared the objectives measured by their test. The qualifications of participants in this aspect of the test development process is important information for potential users.

Table 1

Summary of Ratings of the Criterion-Referenced Tests

Question	Test 1	2	3	4	5	6	7	8	9	10	11
A1	A	A	A	A⁻	A⁻	A	A	A	A	A	X
A2	X	X	X	X	X	X	X	X	X	X	X
A3	A	A	A⁻	A	A	A	A	A	A	A	A
A4	A	A	A	A⁻	A⁻	A	A	A	A	A	X
A5	A⁻	A⁻	A	A	A	X	X	A	A	A	A
A6	A	A	A	A	A	A	A	A	A	A	A
A7	Y	Y	A⁻	Y	Y	Y	A⁻	A	A⁻	A⁻	A⁻
A8	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻
B1	X	X	A	A⁻	A⁻	X	A⁻	Y	A	A	Y
B2	A⁻	A⁻	A	A⁻	A⁻	?[1]	A⁻	A⁻	A	A	A⁻
B3	X	X	X	X	X	X	X	X	X	X	X
B4	A	A	A	A	A	A	A	A	A	A	A
B5	A	A	A	A	A	A	A	A	A	A	A
B6	A	A	A	Y	Y	?	Y	Y	Y	A	Y
B7	A	A	A	A	A	A	A	Y	Y	Y	Y
B8	X	X	A	X	X	X	A⁻	Y	X	X	Y
C1	A	A	A	A	A	?	A	A	A	A	?[2]
C2	A	A	A	A	A	?	A	A	A	A	A
C3	A	A	A	A	A	?	A	A	A	A	A
C4	A	A	A	A	A	?	A	A	A	A	A
D1	A	A	A	A	A	?	A	A	A	A	A
D2	A	A	A	A	A	?	A	A	A	A	A
E1	A⁻	X	A⁻	Y	Y	X	X	Y	X	X	Y
E2	A	A	A	Y	Y	A	A	Y	A	A	Y
E3	A⁻	A⁻	A⁻	A⁻	A⁻	X	X	X	X	X	A⁻
E4	A⁻	A⁻	A⁻	Y	Y	X	X	Y	X	X	Y
F1	A	A	A	Y	A⁻	Y	A	X	A	A	Y
F2	A	A	X	Y	Y	X	X	Y	A	A	Y
F3	A	A	A⁻	Y	Y	Y	A⁻	Y	A⁻	A⁻	Y
G1	A	A	A	X	X	A	A	X	A⁻	A⁻	Y
G2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
H1	A	A	N	N	N	A⁻	A	N	N	N	N
H2	A	A	N	N	N	?	Y	N	N	N	N
H3	A	A	N	N	N	Y	Y	N	N	N	N
I1	A	A	A	A	A	?	A	A	A	A	A
I2	A	A	A	A	A	?	A	A	A	A	A
I3	A	A	A	A	A	?	A	A	A	A	A
J1	A⁻	A⁻	A	Y	Y	?	A⁻	Y	A⁻	A⁻	Y
J2	A	A	A	X	X	?	A	A⁻	A⁻	A⁻	A

[1]We did not have the proper materials to assess the quality of the test in the areas marked by a "?".

[2]The information was on a cassette. We did not listen to the tape and so we were not in a position to rate this aspect of the test.

4. Since test developers have not used "domain specifications", it is impossible to assess "item representativeness". Item representativeness is essential if users desire to use objective scores to "generalize to the domains of behaviors defined by the objectives." If item representativeness is not established, scores can only be interpreted in terms of the specific items included in the test.

5. "Item analysis" is an area in which there are two problems: (a) Too little explanation is offered of the choice of particular item statistics and of the specifics of item statistics usage, and (b) item statistics are used in test construction thereby "biasing" the content validity of the test in unknown ways.

6. Test score reliability was not handled very well in most of the manuals. Either (a) inappropriate information relative to the stated uses of the test scores was offered, or (b) no information was offered.

7. Cut-off scores are typically offered, but there is no rationale offered for setting cut-off scores. Procedures used for setting cut-off scores are not explained, nor is any evidence offered for the "validity" of cut-off scores (for example, do those examinees classified as "masters" typically perform better than "non-masters" on some appropriately chosen external criterion measure?).

8. Factors affecting the validity of scores are not offered in any of the manuals.

9. Only a few of the manuals introduced the notion of "error" in test scores. It is extremely important for users to have some indication of the "stability" of their objective scores and/or "consistency of mastery/non-mastery decisions".

Concluding Remarks

Our proposed guidelines were developed after careful study of the criterion-referenced testing literature and the Test Standards. However, they are offered here only to serve as a "catalyst" for further discussion and debate on a topic of considerable importance to the test and measurement field. Our use of the proposed guidelines to evaluate eleven criterion-referenced tests was intended to (1) demonstrate that the proposed

guidelines were workable, and (2) highlight areas where considerably more (or different) work on the part of test developers is needed.

Our goal for preparing this paper has been accomplished if (1) it stimulates others to extend and improve upon our guidelines, and (2) it helps to direct test developers toward more acceptable practices of criterion-referenced test construction and preparation of test manuals.

Individuals with suggestions for improving the guidelines are encouraged to write the authors.

References

Eignor, D. R. Methodological and psychometric contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1978.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Swezey, R. W., & Pearlstein, R. B. Guidebook for developing criterion-referenced tests. A report prepared for the U.S. Army Research Institute for the Behavioral and Social Sciences. Reston, Virginia: Applied Science Associates, August, 1975.

Walker, C. B. Standards for evaluating criterion-referenced tests. Los Angeles: Center for the Study of Evaluation, UCLA, 1977. (Unpublished manuscript.)

CRITERION-REFERENCED TESTING BIBLIOGRAPHY

(Excerpts from bibliography
Compiled by Ronald K. Hambleton
March 1, 1978)

- Reviews

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1978.

- Test Development

Fremer, J. Handbook for conducting task analyses and developing criterion-referenced tests of language skills. PR 74-12. Princeton, N. J.: Educational Testing Service, 1974.

Jackson, R. Developing criterion-referenced tests. TM Report No. 1. Princeton, N. J.: ERIC Clearing House on Tests, Measurement and Evaluation, 1970.

- Test Reliability

Brennan, R. L., & Kane, M. T. Signal/noise rations for domain-referenced tests. Psychometrika. in press.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-275.

Swaminathan, H., Hambleton, R. K. & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

- Test Validity

Cronbach, L. J. Test validation. In R. L. Thorndike (ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Messick, S. A. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

- Cut-off Scores

Glass, G. V. Standards and Criteria. Journal of Educational Measurement, 1978, 15, in press.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-testing. Review of Educational Research, 1976, 46, 133-158.

- Test Length

Millman, J. Determining test length: Passing scores and test lengths for objectives-based tests. Instructional objectives exchange, Los Angeles, California, 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

- Norms

Popham, W. J. Normative data for criterion-referenced tests? Phi Delta Kappan, 1976, 58, 593-594.

- Applications: General

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. Measurement in Education, 1972, 3, 1-8.

ALTERNATIVE APPROACHES TO EDUCATIONAL ASSESSMENT

National Consortium on Testing
Panel Discussion, October 23, 1978

George E. Hein,
Lesley College

Selected Bibliography on Alternative Approaches in Educational Assessment

(focus on methods and examples with limited discussion of theory)

*Carini, P., "The Child as Reader," and "Documentary Processes," two papers delivered at A.E.R.A., Toronto, April 1978, available from K. McCarthy, V.E.R.B., State Department of Education, Montpelier, VT 05602.

Engel, B.S., Informal Evaluation, Grand Forks, N.D., North Dakota Study Group on Evaluation, 1977.

Guba, Egon, G., Towards a Methodology of Naturalistic Inquiry in Educational Evaluation, Los Angeles: Center for Study of Evaluation, UCLA, 1978.

Hawes, G.R., "Managing Open Education, Testing Evaluation and Accountability," Nation's Schools, 93, 33-47 (1974).

Hein, G.E., "Evaluation in Open Education: Emergence of Qualitative Methodology," forthcoming in S.J. Meisels, editor, Special Education and Development, 1979.

Mehan, H., "Structuring School Structure," Harvard Education Review, 48, 32-64 (1978).

Stake, Robert, Evaluating the Arts in Education, Columbus, Merrill, 1975.

Willis, George, ed., Qualitative Evaluation, Berkeley, McCutchan, 1978.

* * *

Some Examples of Primarily Qualitative (Alternative) Program Evaluations

* Alberty, B., G.A.M.E., Inc., Documentation 1974-75, New York, Game, 1977.

* Aldrich, Ruth Ann, Marcy Open School, 1973-74 Goal Evaluation, Minneapolis: SEA Internal Evaluation Team, 1974.

Chittenden, E., et al. First Year Evaluation Study of the Workshop Center for Open Education, City College of New York, New York: Workshop Center for Open Education, 1973.

* Duckworth, E., The African Primary Science Program: An Evaluation and Extended Thoughts, Grand Forks, North Dakota Study Group, 1978.

* Engel, B.S., "One Way it Can Be," Today's Education, 66, March-April, 1977, pp. 50-52. A summary of an extensive evaluation of third grade achievement at the Cambridge Alternative Public School, 1976.

Engel, B.S. and G.E. Hein, Summary Report of Findings Across Forty School Cultural Institution Integration Programs, Cambridge, Program Evaluation and Research Group, 1977. (Also other reports from this group).

Hanley, J., et al., Curiosity, Competence, Community (An Evaluation of Man: A Course of Study) Cambridge: Education Development Center, 1970.

Tikunoff, W.J., D.C. Berliner and R.C. Rist, An Ethnographic Study of the Forty Classrooms of the Beginning Teacher Evaluation Study Known Sample, San Francisco: Far West Laboratory, 1975.

APGA Policy Statement: Responsibilities of Users of Standardized Tests, from Guidepost, October 5, 1978, pp. 5-8. (ED 177166)

INTRODUCTION

During the past several years, individual APGA members have been under increasing pressure from their various constituencies to define, provide and employ safeguards against the misuse of standardized tests. APGA as an organization has also been challenged by individuals and agencies to provide leadership in the face of growing concern about the effects of testing on clients of all ages and all subpopulations, and in all settings.

At the 1976 APGA convention, the board of directors requested action on the development of a statement on the responsible use of standardized tests. A committee representing all APGA divisions and regions spent two years studying the issues and developing the following statement.

Target audience: The statement is intended to present the position and address the needs of the professional members of APGA divisions and regions. Although this position may provide guidance for test developers, teachers, administrators, parents, press or the general public, it is not designed to represent these audiences. The statement is built on the assumption that test data of themselves are neutral and that guidelines are needed to promote constructive use of tests.

Organization and focus: The statement is organized into eight sections: Introduction, Decision Rules, Test Selection, Qualifications of Test Users, Test Administration, Scoring of Tests, Test Interpretation and Communication. Each section is directed toward the various uses and decisions that must be made by the test user (e.g., whether to test, which test(s) to use, what data to obtain, how to interpret, etc.). The committee developed a classification system for the uses of standardized tests and treated only those issues that fit into the use classification scheme.

The next step was to define issues related to the classification system. Issues were sought from individuals, professional statements, literature and the popular press. Issues were examined in terms of their relevance to APGA members and to their importance in terms of the possible consequences to the person(s) tested.

Only the principles underlying each issue are specified. These principles are appropriate as standards for all APGA divisions and regions. Divisions and regions are encouraged to develop their own statements, expanding on each prrinciple with specific procedures and examples and appropriate to their members. The principles are grouped around similar issues and are indexed for easy reference.

Composition of Committee: Each division president and the representative of each region was asked to appoint to the committee a member who was: (1) knowledgable in the use of standardized tests; (2) aware of the national concerns about the use of tests; (3) willing to involve the division/region membership in identification of the concerns, needs and propositions of the division/region as these relate to the responsible use of tests; (4) able to participate actively and to respond promptly to requests for review of draft documents; (5) capable of securing cooperation of division/region officers in developing procedures for division/region implementation of the statement; (6) willing to accept specific assignments such as providing current organizational statements.

In addition to the division/region appointment, the committee included two members-at-large and a chairperson, all appointed by the APGA president, and three test experts to serve as a core committee to assist the chairperson in analyzing and synthesizing input from committee members and in preparing draft statements for committee review.

To furnish perspective for the work of this committee, a review of relevant literature was conducted and each member of the committee received copies of numerous position papers, reports, articles and monograps that added to understanding the issues and the consequences of alternative principles.

Among these papers were the interim report of the 1975-76 APGA Committee on Standardized Testing; the 1972 APGA/AMEG Statement on the Responsible Use of Tests; position papers of individual divisions and of other professional organizations such as the American Psychological Association, the American Education Research Association and the National Council on Measurement in Education; journal articles; and conference presentations.

The committee's statement is intended to be sensitive to current and emerging problems and concerns that are generic to all APGA divisions/regions and to address these problems and concerns with principles that are specific enough to serve as a template to develop division/region statements addressed to the specific disciplines/settings of individual divisions/regions.

DECISION RULES

In human service agencies, decisions about client needs may be made on the basis of direct observation or historical information alone. Further refinement of direct observation and historical data can often be obtained by employing standardized tests.

Deciding whether to test creates the possibility of three classes of errors relative to the agency functions of description, diagnosis, prescription, selection, placement, prediction, growth evaluation, etc.

First, a decision not to test can result in misjudgments that stem solely from inadequate data.

Second, tests may be used well, producing data that could improve accuracy in decisions affecting the client but that are not utilized.

Third, tests may be misused through inappropriate selection, improper administration, inaccurate scoring, incompetent interpretation or indiscriminate, inadequate, or inaccurate communication.

To reduce the chance for errors, the responsible practitioner will always determine in advance why a given test should be used. This provides protection and benefits for both the client and the agency. Having a clearly developed rationale increases the probable benefits of testing by indicating how a particular set of information, when used by an individual or set of individ-

uals, will contribute to a sounder decision without prejudice to either the client or the agency.

The guidelines that follow are intended to provide decision rules to help agencies and practitioners avoid charge of irresponsible practice.

Defining purposes for testing:

1. Decide whether you will be testing to evaluate individuals, groups or both.

2. Identify your interests in the particular target population in terms of the agency's purposes and capabilities.

3. Determine limits to diagnosis, prediction or selection created by age, racial, sexual, ethnic or cultural characteristics of those to be tested.

4. Develop specific objectives and limits for the use of test data in relation to each of the component service areas of placement/selection, prediction (expectancies), description/diagnosis and growth studies (assessing change over time).

A. Placement: If the purpose is selection or placement (selection is a simple in-out sort of placement), the test selector and interpreter must know about the programs or institutions in which the client may be placed and be able to judge the consequences of such placement or exclusion for the client.

B. Prediction/expectancies: If the purpose is prediction, the persons deciding to test and/or interr ret the results must understand the pitfalls of labeling, stereotyping and prejudging people. Ways to avoid these potentially invidious outcomes should be known.

C. Description/diagnosis: If the purpose is diagnosis or description, the selector or interpreter should understand enough about the general domain being measured to be able to identify those aspects adequately measured and those not.

D. Growth/change assessment: If the purpose is to examine growth, the person designing the study and interpreting the results needs to know the many problems associated with such measurement:

1) the unreliability of change measures;

2) the pitfalls in using norms as reference points;

3) the associated problems of articulation and comparability;

4) the limitations of scoring scales, such as grade equivalents, that may not have the comparable meaning which they appear to have at different scale levels.

Determining information needs:

1. Assess the consequences for the clients of both testing or not testing.

2. Determine what decisions can be made with existing information to avoid unnecessary data-gathering efforts.

3. Limit data gathering to those functions or aptitude, achievement, interests/attitudes/values and perceptual-motor skills that are directly relevant in making decisions about delivery of services to a particular individual or group.

4. Identify whether the test being considered can provide acceptable levels of precision (reliability) for the decision being made.

5. Identify whether the data obtained can be cross-validated against other available data as a part of the decision-making process.

6. Determine the amount and form of data to be shared on the basis of maximum relevance to the agency's purposes and capabilities.

Identifying users of test information:

1. Data should be prepared so that they can be comprehended by the persons using the data for decision-making.

2. Limit access to users specifically authorized by the law or by the client.

3. Identify obsolescence schedules so that stored personal test data may be systematically reclassified and relocated to historical files or destroyed.

4. Process personal data used for research or program evaluation so as to assure individual anonymity.

QUALIFICATIONS OF TEST USERS

All professional personnel and guidance workers should have formal training in psychological and educational measurement and testing. Nevertheless, it is unreasonable to expect that this training necessarily makes one an expert or even that an expert always has all the knowledge and skill appropriate to any particular situation. Thus, questions of user qualifications should always arise when testing is being considered.

Those who participate in any aspect of testing should be qualified to do so. Lack of proper qualifications leads to misuse, errors and sometimes damage to clients. Each professional is responsible for making judgment on this matter in each situation and cannot leave that responsibility either to clients or to those in authority.

In many instances information or skills that may be lacking can be acquired quite readily by those with a background of professional training and experience. In all instances it is incumbent upon the individual to obtain that training or arrange for proper supervision and assistance when engaged in, or planning to engage in, testing.

The requisite qualifications for test users depend on four factors: (1) the particular role of the user: (2) the setting in which the use takes place; (3) the nature of the test; and (4) the purpose of the testing.

These factors interact with each other but may nevertheless be considered separately for the purposes of these standards.

Roles of test users, selectors, administrators, scorers and interpreters:

A test user may play all of these roles or any subset of them when working with other professional personnel. In some situations each role may be the responsibility of a different person. The knowledge and skills that pertain to these roles are listed under the sections so headed. The general principle is that the test users should engage in only those testing activities for which their training and experience qualify them.

Settings and conditions of test use:

Counselors and personnel workers should assess the quality and relevance of their knowledge and skills to the situation before deciding to test or to participate in a testing program.

Characteristics of tests:

Tests differ in many ways, and users need to understand the peculiarities of the instruments they are using.

Purposes of testing:

The purpose of the testing dictates how the test is used and thus may influence requisite qualifications of users beyond those entatiled by their testing roles. Technically proper use for ill-understood purposes may constitute misuse.

TEST SELECTION

Tests should be selected for a specific measurement purpose, use and interpretation. The selection of tests should be guided by information obtained from a careful analysis of the following major considerations:

— What are the characteristics of the population to be tested?

—What knowledge, skills, abilities or attitudes are to be assessed?

— What are the purposes for testing?

— How will the test scores be used and interpreted?

When complete answers to these questions have been obtained, selection or development of tests should be directed toward obtaining measures that are congruent with the stated needs for assessment in terms of the purposes, content, use, interpretation and particular characteristics of the individuals who are to be tested.

Selection of tests must also be guided by the criteria of technical quality recommended by the measurement profession and published by APA/AERA/NCME in "Standards for Educational and Psychological Tests" (1974). Full recognition and analysis of these considerations should become the focus of a process to select appropriate tests. The responsible test selector will:

Select appropriate tests:

1. Select tests that have been demonstrated, to the satisfaction of professional specialists, as appropriate for the characteristics of the population to be tested.
2. Select tests that are within the level of skills of administration and interpretation possessed by the practitioner.
3. Determine whether a common test or different tests are required for the accurate measurement of groups with different characteristics.
4. Recognize that different tests for cultural, ethnic and racial groups constitute inefficient means for making corrections for differences in prior life experiences, except where different languages are involved.
5. Determine whether persons or groups that use different languages should be tested in either or both languages and in some instance by prior testing for bilingualism.

Relate evidence or validity to particular usage:

1. Apply tests or selection only when they show predictive validity for the specific tasks or competencies needed in an educational or employment assignment to maintain legal prescriptions for non-discriminatory practices in selection, employment or placement.
3. Determine validity of a test (whether the test measures what it claims to measure) through evidence of the contructs used in developing the measures, the correlation of the test performance with another appraisal of the characteristics being measured, or the predictions of specified behavior from the test performance.
3. Determine that the content of the test has high congruence with the users' definition of the knowledge and skills that are the desired criteria of human performance to be appraised.
4. Confirm that the criteria of human performance to be appraised are contained in the tasks and results of the testing procedure.

Employ user participation in test selection:

Actively involve the persons who will be using the tests (administering, scoring, summarizing, interpreting, making decisions) in the selection of tests that are congruent with the locally determined purposes, conditions and uses of the measurement.

Select tests to satisfy local use:

1. Give specific attention to how the test is designed to handle the variation of motivation among persons taking the test, the variation or bias in response to the test content and the effects of the presence or absence of guessing in the responses to the test questions.
2. Determine whether tests standardized for nationwide use show evidence that such tests yield comparable results for individuals or groups with cultural differences.
3. Identify and analyze the effects of working speed and language facility in relation to the criteria of human performance that are expected to result from the test.

Consider technical characteristics of tests:

1. Select only published or locally developed tests that have documented evidence of the reliability or consistency of the measure.
2. Select tests that have documented evidence of the effectiveness of the measure for the purpose to be served: placement/selection, prediction (expectancy), description/diagnosis, or growth studies (change over time). A test is rarely equally effective for the four common test uses.
3. Consider the procedures used in standardization and norming for relevance to the local population and the desired use and interpretation.
4. Use separate norms for men and women only when empirical evidence indicates this is necessary to minimize bias.
5. Determine the degree of reliability (or validity) demanded of a test on the basis of the nature of the decisions to be based on test scores.
6. A test for final diagnosis or selection requires a higher degree of reliability than an initial screening test.
7. Explicitly list and use the ease and accuracy of the procedures for scoring, summarizing and communicating test performance as criteria for selecting a test.
8. Recognize that the technical characteristics and norms of standardized tests may vary when used with different populations. The selection process should include trial administrations to verify that the test is functioning with the technical characteristics and desired results for the local population and local uses.

Practical constraints of cost, conditions and time for testing must be considered but not used as the primary criteria for test selection.

TEST ADMINISTRATION

Test administration includes all procedures that are used to ensure that the test is presented consistently in the manner specified by the test developers and used in the standardization and that the individuals being tested have orientation and conditions that maximize opportunity for optimum performance.

Standardized tests should provide manuals giving specific directions for administering, scoring and interpreting tests. Tests developed for a specific local purpose, use or population should be administered in a prescribed and consistent manner to obtain optimum

performance from the individuals being tested. Effective administration of tests requires that the administrator have knowledge and training with the instruments and the processes of presentation.

Orientation:

1. Inform testing candidates, relevant institutions or agencies and the community about the testing procedure. The orientation should describe the purposes and contents sampled by the test, how it is administered and how the scores will be reported and used.

2. Provide annual training for test administrators by qualified professional specialists if your agency or institution uses tests or sponsors testing programs.

3. Routinely review the test materials and administration conditions well in advance of the time for testing so that full preparation will ensure standardized administration and recognition of any irregularities that may occur.

4. Ensure that the orientation is sufficient to make the test relevant for the individual or group being tested before beginning test administration.

5. Ensure that all persons being tested have the specified practice with sample problems or test taking skills prior to their performance on the test.

6. Demonstrate the techniques and requirements for marking machine-scorable answer sheets. Check all individuals taking a test for competency in the techniques of recording their answers prior to the specific period of testing.

Qualifications of test administrators:

1. Administer standardized tests only if you are qualified by training and experience as competent to administer particular tests.

2. Know the exact population and procedures used in standardizing the test and determine that the test is appropriate for the local population that is to be tested.

3. Acquire extensive training required to administer, score or interpret tests requiring test-specific training.

Giving directions:

1. Administer standardized tests with the verbatim instructions, exact sequence and timing and the identical materials that were used in the test standardization.

2. Present all tests (whether standardized, published or locally constructed) in an identical manner to ensure that the test is a fair and comparable demonstration of the performance of each individual taking the test.

Recognize that taking a test may be a new and frightening experience or stimulate anxiety or frustration for some individuals. Communicate to the examinees that they should attempt each task with positive application of their skills and knowledge and the anticipation that they will do their best.

Testing conditions:

1. Devote concentrated attention to observing the condition and reactions of the individuals being tested. Observe those being tested and identify environmental, health or emotional conditions that should be recorded and considered as invalidating elements for the test performance.

2. Possess and demonstrate clear verbal articulation, calmness and positive anticipation, empathy for and social identification with the examiness, and, impartial treatment for all being tested.

3. Determine whether the testing environment (seating, work surfaces, lighting, heating, freedom from distractions, etc.) is conducive to the best possible performance of the test-takers.

4. Administer tests in physical facilities and psychological climates that allow each individual being tested to achieve optimum performance.

5. Record any deviation from standardized test administration procedure (such as used to accommodate handicapping conditions) and make it a permanent attachment to the test score or record.

6. Develop and complete for each test a systematic and objective procedure for observing and recording the behavior of those being tested (and recording conventional or deviant conditions of testing). Attach this record to the test scores of the persons tested.

7. Provide a written record of any circumstances that may have increased or reduced the opportunity of an individual being tested to demonstrate his or her best performance.

Accept responsibility for seeing that invalid or questionable test scores are not recorded, or not recorded without written qualification of the conditions that may have affected optimum test performance.

9. Arrange assistance from trained personnel in providing uniform conditions and in observing the conduct of the examinees when large groups of individuals must be tested.

Professional collaboration:

Recognize that in institutional settings, and wherever skill and knowledge can be pooled and responsibility shared, it is the qualifications of the team as a whole that count rather than those of individuals. However, coordination and consistency must be maintained.

TEST SCORING

The measurement of human performance depends on accurate and consistent application of defined procedures for crediting the responses made by persons being tested. The procedures for scoring and recording test performance must be continuously audited for consistency and accuracy.

1. Routinely rescore a sample of the test answer sheets to verify the accuracy of the initial scoring.

2. Employ systematic procedures to verify the accuracy and consistency of machine scoring of answer sheets.

3. Obtain a separate and independent verification that appropriate scoring rules and normative conversions are used for each person tested.

4. Verify as accurate the computation of raw scores and the conversion of raw scores to normative or descriptive scales prior to release of such information to the tested person or to users of the test results.

5. Routinely check machine or manual reports of test results for accuracy. The person performing this task must be qualified to recognize inappropriate or impossible scores.

6. Develop and use systematic and objective procedures for observing and recording the conditions and behaviors of persons being tested and make this a part of the scores or test results that are reported.

7. Clearly label the scores that are reported and the date that a particular test was administered.

TEST INTERPRETATION

Test interpretation encompasses all the ways we assign value to the scores.

A test can be described as a systematic set or series of standard observations of performances that all fall in some particular domain. Typically each observation yields a rating of the performance (such as right or wrong and pass or fail), then these ratings are counted and this count becomes the basis of the scores. Such scores are usually much more stable than the result of any single performance. This score reliability creates the possibility of validity greater than can be obtained from unsystematic or nonaggregated observations.

The proper interpretation of test scores starts with understanding these fundamental characteristics of tests. Given this, the interpretation of scores from a test entails knowledge about (1) administration and scoring procedures; (2) scores, norms, and related technical features; (3) reliability; and (4) validity.

Adequate test interpretation requires knowledge and skill in each of these areas. Some of this information can be mastered only by studying the manual and other materials of the test; no one should undertake the interpretation of scores on any test without such study.

Administration and Scoring:

Standard procedures for administering and scoring the test limit the possible meanings of scores. Departures from standard conditions and procedures modify and often invalidate the criteria for score interpretation.

1. The principles in the section on administration and scoring need to be understood by those engaged in interpretation.

2. Ascertain the circumstances peculiar to the particular administration and scoring of the test.

A. Examine all reports from administrators, proctors and scorers concerning irregularities or conditions, such as excessive anxiety, which may have affected performance.

B. Weigh the possible effects on test scores of examiner-examinee differences in ethnic and cultural background, attitudes and values in light of research on these matters. Recognize that such effects are probably larger in individual testing situations.

c. Look for administrators' reports of examinee behavior that indicate the responses were made on some basis other than that intended — as when a student being tested for knowledge of addition-number-facts adds by making tallies and then counting them.

3. Consider differences among clients in their reaction to instructions about guessing and scoring.

4. Recognize or judge the effect of scorer biases and judgment when subjective elements enter into scoring.

Scores, norms and related technical features:

The result of scoring a test is usually a number (or a set of numbers) called a raw score. Raw scores taken by themselves are not usually interpretable. Some additional steps must be taken.

The procedures either translate the numbers directly into descriptions of their meaning (e.g., pass or fail) or into other numbers called derived scores (e.g., standard scores) whose meaning stems from the test norms.

To interpret test scores, these procedures and the resulting descriptions or derived scores need to be thoroughly understood. Anything less than full understanding is likely to produce at least some, and probably many, serious errors in interpretation. The following are imperatives for interpreting tests:

1. Examine the test manuals, handbooks, users' guides and technical reports to determine what descriptions or derived scores are produced and what unique characteristics each may have.

2. Recognize that direct score interpretations such as mastery and nonmastery in criterion-referenced tests depend on arbitrary rules or standards.

A. Report number or percent of items right in addition to the indicated interpretation whenever it will help others understand the quality of the examinee's test performance.

B. Recognize that the difficulty of a fixed standard, such as 80 percent right, will vary widely from objective objective. Such scores are not comparable in the normative sense.

C. Recognize that when each score is classified as pass fail, mastery-nonmastery or the like, that each element is being given equal weight.

3. Use the derived scores that fit the needs of the current use of the test.

A. Use percentile ranks for direct comparison of individuals to the norm or reference group.

B. Use standard scores or equal unit scaled scores whenever means and variances are calculated or other arithmetic operations are being used.

4. Recognize that only those derived scores that are based on the same norm group can be compared.

5. Consider the effect of any differences between the tests in what they measure when one test or form is equated with another, as well as the errors stemming from the equating itself.

Give greater credence to growth or change shown by the same test (including level and form) than to equated measures except where practice effects or feedback have destroyed the validity of a second use.

6. Evaluate the appropriateness of the norm groups available as bases for interpreting the scores of clients.

A. Use the norms for the group to which the client belongs.

B. Consider using local norms and derived scores based on these local norms whenever possible.

7. Acquire knowledge of specific psychological or educational concepts and theories before interpreting the scores of tests based on such knowledge.

Reliability:

Reliability is a prerequisite to validity. Generally, the greater the number of items the greater the reliability of the test. The degree to which a score or a set of scores may vary because of measurement error is a central factor in interpretation.

1. Use the standard error of measurement to obtain a rough estimate of the probable variation in scores due to unreliability.

2. Use the reliability coefficient to estimate the proportion of score variance that is not due to error.

3. Consider the sources of variance attributable to error in the particular reliability indexes reported in relationship to the uses being made of the scores.

4. Assess reported reliabilities in light of the many extraneous factors that may have artificially raised or lowered these estimates, such as test speededness, sample homogeneity or heterogeneity, restrictions in range and the like.

5. Distinguish indexes of rater reliability (i.e., of objectivity) from test reliability.

Validity:

Proper test interpretation requires knowledge of the validity evidence available for the test as used. Its validity for other uses is not relevant. The purpose of testing dictates how a test is used. Technically proper use for ill-understood purposes may constitute misuse. The nature of the validity evidence required for a test is a function of its use.

Prediction — developing expectancies: The relationship of the test scores to an independently developed criterion measure is the basis for predictive validity.

1. Consider both the reliability and the relevance of the criterion measures used.

2. Use cross validation data to judge the validity of predictions.

3. Question the meaning of an apparently valid predictor that lacks both construct and content validity. Assess the role of underlying and concomitant variables.

4. Consider the validity of a given measure in the context of all the predictors used or available. Does the measure make an independent contribution to the prediction over and above that provided by other measures?

5. Consider the pitfalls of labeling, stereotyping and prejudging people. The self-fulfilling prophecies that may result are often undesirable.

Placement/selection: Predictive validity is the usual basis for valid placement. Consider the evidence of validity for each alternative (i.e., each placement) when inferring the meaning of scores.

1. Obtain adequate information about the programs or institutions in which the client may be placed in order to judge the consequences of such placement.

2. Estimate the probability of favorable outcomes for each possible placement (e.g., both selection and rejection) before judging the import of the scores.

3. Consider the possibility that outcomes favorable from an institutional point of view may differ from those that are favorable from the examinees' point of view.

4. Examine the possibility that the clients' group membership (race, sex, etc.) may alter the reported validity relationships.

5. Use all the available evidence about the individual to infer the validity of the score for that individual. Each single piece of information about an individual, (e.g, test score, teacher report, or counselor opinion) improves the probability that proper judgments and decisions can be made.

A. Test scores should be considered in context; they do not have absolute meaning.

B. Single test scores should not be the sole basis for placement or selection.

Description/diagnosis: Distinguish between those descriptions and diagnoses using psychological constructs that can be validated only indirectly and those for which content specifications suffice.

1. Identify clearly the domain specified by those asserting content validity. Assess the adequacy of the content sampling procedures used in writing and selecting items.

2. Identify the dimensions of the construct being measured when multiple scores from a battery or inventory are used for description.

A. Examine the content validity and/or the construct validity of each score separately.

B. Consider the relative importance of the various subtests, parts, objectives or elements yielding separate scores and judge the weight they each should be given in interpretation.

C. Recognize that when scores are summed or averaged their weight is a function of their variances.

D. Recognize that when each score is classified as pass-fail, mastery-nonmastery or the like, each element is being given equal weight.

3. Examine the completeness of the description provided, recognizing that no set of test scores completely describes a human being.

Growth — Studies of change: Valid assessment of growth or change requires both a test having descriptive validity and a procedure for establishing that the scores obtained differ from those that might arise when no change has occurred.

1. Report as possibilities all the interpretations the study or evaluation design permits. Point out those interpretations that are precluded by the design used.

A. When standard procedures such as the RMC models (see Tallmadge, GK. and Horst, D.P., A procedural guide for validating achievement gains in educational projects; Mountain View, Calif: RMC Research Corp., Dec. 1975) for measuring growth in achievement are employed, use the descriptions of strengths and weaknesses provided.

B. Look for naturally occurring control groups not part of design whenever possible.

2. Consider the strengths and weaknesses of the particular tests used with respect to this use.

Consider the possibility of floor or ceiling effects, the content changes level to level, the adequacy of articulation in multilevel tests, the comparability of alternate forms, the adequacy of the score-equating across forms, and the comparability of timing of the testing to that of the norming.

3. Recognize the unreliability of individual score differences as measures of change.

4. Recognize the limitations of scoring scales such as grade equivalents that may not have the comparable meaning they appear to have at different levels of the scale.

5. Recognize the need for equal interval scales when trying to assess the amount of change.

COMMUNICATING TEST RESULTS

Communication consists of reporting data in such a way that it is comprehensible and informative. The responsible practitioner reports test data with a concern for the user's need for information and the purposes of evaluating the significance of the information.

There must also be a concern for the right of the individual tested to be informed regarding how the results will be used for his or her benefit (informed consent), who will have access to the results (right to

privacy), and what safeguards exist to prevent misuse.

Where standardized test data are being used to enhance decisions about an individual, the practitioner's responsibilities are as follows:

Know the manual:

1. Become thoroughly familiar with the publisher's manual before attempting to "explain" any results.
2. Develop skills needed to communicate results of tests, using concepts that are frequently misunderstood before communicating results to clients, the public, or other recipients of the information.

Know the limits:

1. Inform the person receiving the test information that "scores" are approximations, not absolutes, and indicate the SEM or the margin of error in some other way, such as by reporting score intervals rather than points.
2. Candidly discuss with the person receiving the test information any qualifications necessary to understand potential sources of bias for a given set of test results relative to their use with a specific individual;
3. Emphasize that test data represent just one source of information and should rarely, if ever, be used alone for decision making.

Informed consent:

1. Inform the person receiving the test information of any circumstances that could have affected the validity or reliability of the results.
2. Inform the examinee of what action will be taken by the agency and who will be using the results.
3. Obtain the consent of the examinee before using test results for any purpose other than that advanced prior to testing.

Right to privacy:

Inform the examinee of steps to be taken to correct any erroneous information that may be on file as a result of testing.

Where standardized test data are being used to describe **groups** for the purpose of evaluation, the practitioner's responsibilities are as follows:

Background information:

1. Include background information to improve the accuracy of understanding about any numerical data.
2. Identify the purposes for which the reported data would be appropriate.

Politics:

Be aware that public release of test information provides data for all kinds of purposes and that some of these may be adverse to the interests of those tested.

Averages and norms:

1. Clarify in particular that "average" on a standardized test is a range, not a point, and typically includes the middle 50 percent of the group being considered.
2. Qualify all group data in terms of the appropriateness of the norms for that group.

Agency policies:

1. Work for agency test-reporting policies designed to strengthen and protect the benefits of the groups being measured.
2. Work within the agency to establish procedures for periodic review of internal test use.

ANNOUNCEMENT

The Committee on Ability Testing of the National Research Council will hold hearings on November 17-18, 1978 at the National Academy of Sciences in Washington, D.C. As part of its continuing examination of the needs served and the issues generated by the use of standardized tests in educational and employment settings, the Committee wants to provide interested groups and individuals the opportunity to present their views and supply pertinent supporting data.

Given the large number of organizations interested in testing issues and the fact that the time available will not permit the Committee to receive the oral testimony of all appropriate persons, the Committee will select among those offering to participate in order to insure that a full range of positions will be represented at the meeting. All others will be encouraged to submit written statements. It is anticipated that all statements will become part of the official record of the meeting.

The meeting will be held from 9:00 to 12:00 each morning and from 2:00 to 4:30 in the afternoon. Members of the public will be welcome to attend.

For further information contact:

Alexandra K. Wigdor
Committee on Ability Testing
National Academy of Sciences, JH-813
2101 Constitution Ave., N.W.
Washington, D.C. 20418
(202)389-6951

September, 1978

National Academy of Sciences - National Research Council

Assembly of Behavioral and Social Sciences

COMMITTEE ON ABILITY TESTING

The Committee on Ability Testing was established against a background of widespread public concern about the role of testing in American life. Many of the most important decisions made about an individual--placement in advanced, remedial, or special education programs in primary and secondary schools; admission to college and professional schools; selection for apprenticeship, employment, and military assignment--are made, at least in part, on the basis of test scores.

Because testing affects so many people in important ways, debate has occurred on many fronts. Some groups are enthusiastic about the potential of tests for furthering merit selection, revealing hitherto undiscovered intellectual talents, and increasing efficiency and accountability in a variety of educational and employment settings. Others have found the negative effects of test use and test abuse more compelling. Spokesmen for minority interests, for example, have attacked standardized tests as barriers to social equality, a concern that has been reiterated in the Title VII case law on employment discrimination. Both high expectations and serious complaints have focused public attention on the underlying questions of what tests actually measure and the meaning to be attached to test scores. The increasing interest of courts, legislatures, and governmental agencies since the passage of the Civil Rights Act of 1964 in the use of tests as selection devices has added a significant new dimension to these questions.

The Committee on Ability Testing, which is currently receiving support from the Carnegie Corporation, the Ittleson Foundation, the National Institute of Education, the National Institute of Mental Health, and the Civil Service Commission, has undertaken a major examination of the role of standardized testing in American society. Tests of aptitude, intelligence, achievement, or ability that are used to make decisions about individuals or policies are being considered.

In order to assess accurately the needs served by testing and the problems generated by test use, the Committee is compiling data on the nature, incidence, and impact of testing in educational and employment settings. Close attention is being given to the legal, political, and social contexts within which testing takes place. Because the Committee is a large one with an especially broad mandate, the members have decided to form four subcommittees, two dealing with educational testing and two with employment testing.

It is expected that the Committee's work will result in a report that illuminates the problems posed by the use of standardized tests and recommends appropriate policy alternatives.

A list of Committee members and staff is attached.

COMMITTEE ON ABILITY TESTING

*Wendell R. Garner (Chairman)
Dean of the Graduate School
Yale University
(psychology)

Marcus Alexis
Department of Economics
Northwestern University

William Bevan
Department of Psychology
Duke University

*Lee J. Cronbach
School of Education
Stanford University
(educational psychology, psychometrics)

*Zvi Griliches
Department of Economics
Harvard University

Oscar Handlin
History Department
Harvard University

Delmos Jones
Anthropology Department
City University of New York

Lyle V. Jones
Dean of the Graduate School
University of North Carolina
(psychometrics, statistics)

Philip B. Kurland
School of Law
University of Chicago

Burke Marshall
School of Law
Yale University

William James McGill
President
Columbia University
(psychology)

Melvin R. Novick
Lindquist Center for Measurement
University of Iowa
(statistics, psychometrics)

William Raspberry
The Washington Post
(journalism)

Lauren B. Resnick
Learning Research and Development Center
University of Pittsburgh
(psychology, educational psychology)

Alice Rossi
Department of Sociology
University of Massachusetts

*William Sewell
Department of Sociology
University of Wisconsin-Madison

Janet T. Spence
Department of Psychology
University of Texas

Alan A. Stone
School of Law
Harvard University
(psychiatry, law)

Mary L. Tenopyr
Human Resources Laboratory
American Telephone and Telegraph Company
(industrial psychology)

*John W. Tukey
Bell Telephone Laboratories, Inc.
(statistics)

E. Belvin Williams
Educational Testing Service
(psychology)

Alexandra Wigdor, Acting Study Director
Deborah Coates, Research Associate
Susan S. Sherman, Senior Research Associate

*Member of the National Academy of Sciences

OUTLINE FOR DESCRIPTIVE REVIEW

OF COMMON STANDARDIZED TEST SERIES

Test Title

Author(s)

Publisher

Copyright date(s)

Price

I. General Information

- Type of test (general area tested, such as reading or math; achievement or ability; group or individual test)
- Intended population (age or grade-range, characteristics)
- Content (subtests)
- Forms (number and whether they are presumed equivalent)
- Type of items (multiple choice, true-false, short or extended answers)
- Time required for administration
- Scoring (scales used, scores derived)
- Summary of levels and subtests (in chart form)

II. Theory, Construction and Intended Use

- Development (when test was first developed, and editions through which it has evolved, relationship of current edition to previous ones)
- Rationale for choice of content and skill areas (specifications used in item writing; procedures and results of item analysis)
- Uses of tests (intended uses, potential misuses, other information to consider when interpreting test scores, qualifications of test users)

III. Practical Considerations

- Materials (manuals, guides, test booklets for teachers, students and others)

- Administration of test (setting, timing, directions and sample questions)

- Scoring arrangements (machine or hand-scored)

IV. Technical Quality

- Norms and Scales (norming populations and subgroups identified; method of sampling for norm groups; possible biases, reasons and recommendations for different scales)

- Validity (Predictive, concurrent, content and construct; for each test and subtest, and subgroups of intended populations; and means of determination of different forms of validity evidence)

- Reliability (standard error of measurement, comparability of forms, internal consistency, stability over time in terms of correlation coefficients or components of variance)

V. Outside Reviews

- In Buros' Mental Measurements Yearbook

- By CSE Test Evaluation project

VI. Comments

VII. Sources